Lead Scoring in X Education

The company wishes to identify the most potential leads, also known as 'Hot Leads'.

If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goals of the Case Study

There are quite a few goals for this case study:

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Results Expected

A well-commented Jupyter notebook with at least the logistic regression model, the conversion predictions and evaluation metrics. The word document filled with solutions to all the problems. The overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly Explain the results in business terms Include visualisations and summarise the most important results in the presentation A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

You need to submit the following four components:

- Python commented file: Should include detailed comments and should not contain unnecessary pieces of code.
- Word File: Answer all the questions asked by the company in the word document provided.
- Presentation: Make a presentation to present your analysis to the chief data scientist of your company (and thus you should include both technical and business aspects). The presentation should be concise, clear, and to the point. Submit the presentation after converting it into PDF format.

- PDF File: Write the summary report in a word file and submit it as a PDF.

Model Predict a Customer is a HotLead or not

```
# Suppressing Warnings
import warnings
warnings.filterwarnings('ignore')

# Importing Pandas and NumPy
import pandas as pd, numpy as np

# For spliting train & test
from sklearn.model_selection import train_test_split

# For scaling Features
from sklearn.preprocessing import StandardScaler

# Importing matplotlib and seaborn
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

# For model building
import statsmodels.api as sm

# Feature Selection using RFE
from sklearn.linear_model import LogisticRegression
from sklearn.feature_selection import RFE
```

Load & Understand Data

```
leads = pd.read_csv('Leads.csv')
leads.head()
```

```
                            Prospect ID  Lead Number                Lead
Origin  \
0  7927b2df-8bba-4d29-b9a2-b6e0beafe620       660737
API
1  2a272436-5132-4136-86fa-dcc88c88f482       660728
API
2  8cc8c611-a219-4f35-ad23-fdfd2656bd8a       660727  Landing Page
Submission
3  0cc2df48-7cf4-4e39-9de9-19797f9b38cc       660719  Landing Page
Submission
4  3256f628-e534-4826-9d63-4a8b88782852       660681  Landing Page
Submission

      Lead Source Do Not Email Do Not Call  Converted  TotalVisits  \
0      Olark Chat           No          No          0          0.0
```

```
1  Organic Search            No          No          0          5.0
2  Direct Traffic            No          No          1          2.0
3  Direct Traffic            No          No          0          1.0
4          Google            No          No          1          2.0

   Total Time Spent on Website  Page Views Per Visit  ...  \
0                            0                   0.0  ...
1                          674                   2.5  ...
2                         1532                   2.0  ...
3                          305                   1.0  ...
4                         1428                   1.0  ...

  Get updates on DM Content    Lead Profile    City  \
0                        No          Select  Select
1                        No          Select  Select
2                        No  Potential Lead  Mumbai
3                        No          Select  Mumbai
4                        No          Select  Mumbai

  Asymmetrique Activity Index Asymmetrique Profile Index  \
0                  02.Medium                  02.Medium
1                  02.Medium                  02.Medium
2                  02.Medium                    01.High
3                  02.Medium                    01.High
4                  02.Medium                    01.High

  Asymmetrique Activity Score Asymmetrique Profile Score  \
0                        15.0                        15.0
1                        15.0                        15.0
2                        14.0                        20.0
3                        13.0                        17.0
4                        15.0                        18.0

  I agree to pay the amount through cheque  \
0                                       No
1                                       No
2                                       No
3                                       No
4                                       No

  A free copy of Mastering The Interview Last Notable Activity
0                                     No              Modified
1                                     No          Email Opened
2                                    Yes          Email Opened
3                                     No              Modified
4                                     No              Modified

[5 rows x 37 columns]
```

```
leads.shape
```

```
(9240, 37)

leads.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count
Dtype
---  ------                                         --------------
-----
 0   Prospect ID                                    9240 non-null
object
 1   Lead Number                                    9240 non-null
int64
 2   Lead Origin                                    9240 non-null
object
 3   Lead Source                                    9204 non-null
object
 4   Do Not Email                                   9240 non-null
object
 5   Do Not Call                                    9240 non-null
object
 6   Converted                                      9240 non-null
int64
 7   TotalVisits                                    9103 non-null
float64
 8   Total Time Spent on Website                    9240 non-null
int64
 9   Page Views Per Visit                           9103 non-null
float64
 10  Last Activity                                  9137 non-null
object
 11  Country                                        6779 non-null
object
 12  Specialization                                 7802 non-null
object
 13  How did you hear about X Education             7033 non-null
object
 14  What is your current occupation                6550 non-null
object
 15  What matters most to you in choosing a course  6531 non-null
object
 16  Search                                         9240 non-null
object
 17  Magazine                                       9240 non-null
object
 18  Newspaper Article                              9240 non-null
object
 19  X Education Forums                             9240 non-null
```

```
object
 20  Newspaper                                      9240 non-null
object
 21  Digital Advertisement                          9240 non-null
object
 22  Through Recommendations                        9240 non-null
object
 23  Receive More Updates About Our Courses         9240 non-null
object
 24  Tags                                           5887 non-null
object
 25  Lead Quality                                   4473 non-null
object
 26  Update me on Supply Chain Content              9240 non-null
object
 27  Get updates on DM Content                      9240 non-null
object
 28  Lead Profile                                   6531 non-null
object
 29  City                                           7820 non-null
object
 30  Asymmetrique Activity Index                    5022 non-null
object
 31  Asymmetrique Profile Index                     5022 non-null
object
 32  Asymmetrique Activity Score                    5022 non-null
float64
 33  Asymmetrique Profile Score                     5022 non-null
float64
 34  I agree to pay the amount through cheque       9240 non-null
object
 35  A free copy of Mastering The Interview         9240 non-null
object
 36  Last Notable Activity                          9240 non-null
object
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB

leads.describe()

         Lead Number    Converted  TotalVisits  Total Time Spent on
Website  \
count    9240.000000  9240.000000  9103.000000
9240.000000
mean   617188.435606     0.385390     3.445238
487.698268
std     23405.995698     0.486714     4.854853
548.021466
min    579533.000000     0.000000     0.000000
0.000000
```

```
25%    596484.500000     0.000000     1.000000
12.000000
50%    615479.000000     0.000000     3.000000
248.000000
75%    637387.250000     1.000000     5.000000
936.000000
max    660737.000000     1.000000   251.000000
2272.000000
```

```
       Page Views Per Visit  Asymmetrique Activity Score  \
count           9103.000000                  5022.000000
mean               2.362820                    14.306252
std                2.161418                     1.386694
min                0.000000                     7.000000
25%                1.000000                    14.000000
50%                2.000000                    14.000000
75%                3.000000                    15.000000
max               55.000000                    18.000000
```

```
       Asymmetrique Profile Score
count                 5022.000000
mean                    16.344883
std                      1.811395
min                     11.000000
25%                     15.000000
50%                     16.000000
75%                     18.000000
max                     20.000000
```

Clean & Prepare Data

```
# Check if id fields have duplicated value
# Function return True if no duplicated data in a column
sum(leads.duplicated(subset='Prospect ID'))
```

```
0
```

```
sum(leads.duplicated(subset='Lead Number'))
```

```
0
```

=> No duplicated base on checking the id fields

List of dropped columns

```
dropped_columns = ['Prospect ID', 'Lead Number']
```

Handle NULL & non-sense values

```
# NULL Values Pecentage
round(leads.isnull().sum()/len(leads.index)*100, 2)
```

```
Prospect ID                                    0.00
Lead Number                                    0.00
Lead Origin                                    0.00
Lead Source                                    0.39
Do Not Email                                   0.00
Do Not Call                                    0.00
Converted                                      0.00
TotalVisits                                    1.48
Total Time Spent on Website                    0.00
Page Views Per Visit                           1.48
Last Activity                                  1.11
Country                                       26.63
Specialization                                15.56
How did you hear about X Education            23.89
What is your current occupation               29.11
What matters most to you in choosing a course 29.32
Search                                         0.00
Magazine                                       0.00
Newspaper Article                              0.00
X Education Forums                             0.00
Newspaper                                      0.00
Digital Advertisement                          0.00
Through Recommendations                        0.00
Receive More Updates About Our Courses         0.00
Tags                                          36.29
Lead Quality                                  51.59
Update me on Supply Chain Content              0.00
Get updates on DM Content                      0.00
Lead Profile                                  29.32
City                                          15.37
Asymmetrique Activity Index                   45.65
Asymmetrique Profile Index                    45.65
Asymmetrique Activity Score                   45.65
Asymmetrique Profile Score                    45.65
I agree to pay the amount through cheque       0.00
A free copy of Mastering The Interview         0.00
Last Notable Activity                          0.00
dtype: float64
```

Add columns which have more than 45% of null values into dropped_columns

```
null_columns = ['Lead Quality', 'Asymmetrique Activity Index',
'Asymmetrique Profile Index', 'Asymmetrique Activity Score',
```

```
'Asymmetrique Profile Score']
dropped_columns = dropped_columns + null_columns
dropped_columns
```

```
['Prospect ID',
 'Lead Number',
 'Lead Quality',
 'Asymmetrique Activity Index',
 'Asymmetrique Profile Index',
 'Asymmetrique Activity Score',
 'Asymmetrique Profile Score']
```

```
# drop columns
leads.drop(dropped_columns, 1, inplace=True)
leads.head()
```

```
                Lead Origin     Lead Source Do Not Email Do Not Call  \
0                       API      Olark Chat           No          No
1                       API  Organic Search           No          No
2  Landing Page Submission  Direct Traffic           No          No
3  Landing Page Submission  Direct Traffic           No          No
4  Landing Page Submission          Google           No          No

   Converted  TotalVisits  Total Time Spent on Website  Page Views Per
Visit  \
0          0          0.0                            0
0.0
1          0          5.0                          674
2.5
2          1          2.0                         1532
2.0
3          0          1.0                          305
1.0
4          1          2.0                         1428
1.0

             Last Activity Country  ... Through Recommendations  \
0  Page Visited on Website     NaN  ...                       No
1             Email Opened   India  ...                       No
2             Email Opened   India  ...                       No
3              Unreachable   India  ...                       No
4        Converted to Lead   India  ...                       No

  Receive More Updates About Our Courses
Tags  \
0                                     No          Interested in other
courses
1                                     No
Ringing
2                                     No  Will revert after reading
```

```
the email
3                                       No
Ringing
4                                       No  Will revert after reading
the email

  Update me on Supply Chain Content Get updates on DM Content    Lead
Profile  \
0                                   No                          No
Select
1                                   No                          No
Select
2                                   No                          No
Potential Lead
3                                   No                          No
Select
4                                   No                          No
Select

     City I agree to pay the amount through cheque  \
0  Select                                         No
1  Select                                         No
2  Mumbai                                         No
3  Mumbai                                         No
4  Mumbai                                         No

  A free copy of Mastering The Interview Last Notable Activity
0                                     No              Modified
1                                     No          Email Opened
2                                    Yes          Email Opened
3                                     No              Modified
4                                     No              Modified

[5 rows x 30 columns]
```

Treatment for Select Value as NaN

```
leads = leads.replace('Select', np.nan)



leads.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 30 columns):
 #   Column                                         Non-Null Count
Dtype
---  ------                                         --------------
-----
```

```
 0   Lead Origin                                    9240 non-null
object
 1   Lead Source                                    9204 non-null
object
 2   Do Not Email                                   9240 non-null
object
 3   Do Not Call                                    9240 non-null
object
 4   Converted                                      9240 non-null
int64
 5   TotalVisits                                    9103 non-null
float64
 6   Total Time Spent on Website                    9240 non-null
int64
 7   Page Views Per Visit                           9103 non-null
float64
 8   Last Activity                                  9137 non-null
object
 9   Country                                        6779 non-null
object
 10  Specialization                                 5860 non-null
object
 11  How did you hear about X Education             1990 non-null
object
 12  What is your current occupation                6550 non-null
object
 13  What matters most to you in choosing a course  6531 non-null
object
 14  Search                                         9240 non-null
object
 15  Magazine                                       9240 non-null
object
 16  Newspaper Article                              9240 non-null
object
 17  X Education Forums                             9240 non-null
object
 18  Newspaper                                      9240 non-null
object
 19  Digital Advertisement                          9240 non-null
object
 20  Through Recommendations                        9240 non-null
object
 21  Receive More Updates About Our Courses         9240 non-null
object
 22  Tags                                           5887 non-null
object
 23  Update me on Supply Chain Content              9240 non-null
object
 24  Get updates on DM Content                      9240 non-null
```

```
object
 25  Lead Profile                                   2385 non-null
object
 26  City                                           5571 non-null
object
 27  I agree to pay the amount through cheque       9240 non-null
object
 28  A free copy of Mastering The Interview         9240 non-null
object
 29  Last Notable Activity                          9240 non-null
object
dtypes: float64(2), int64(2), object(26)
memory usage: 2.1+ MB
```

Value Counts Categorical Columns

```
leads['Lead Origin'].value_counts(dropna=False)
```

```
Landing Page Submission    4886
API                        3580
Lead Add Form               718
Lead Import                  55
Quick Add Form                1
Name: Lead Origin, dtype: int64
```

```
leads['Lead Source'].value_counts(dropna=False)
```

```
Google               2868
Direct Traffic       2543
Olark Chat           1755
Organic Search       1154
Reference             534
Welingak Website      142
Referral Sites        125
Facebook               55
NaN                    36
bing                    6
google                  5
Click2call              4
Press_Release           2
Social Media            2
Live Chat               2
youtubechannel          1
testone                 1
Pay per Click Ads       1
welearnblog_Home        1
WeLearn                 1
blog                    1
```

```
NC_EDM                    1
Name: Lead Source, dtype: int64
```

```
leads['Do Not Email'].value_counts() # Yes: 0.08% over total rows => drop
```

```
No     8506
Yes     734
Name: Do Not Email, dtype: int64
```

```
leads['Do Not Call'].value_counts() # Yes 2/9240 => Drop
```

```
No     9238
Yes       2
Name: Do Not Call, dtype: int64
```

```
leads['Last Activity'].value_counts(dropna=False)
```

```
Email Opened                    3437
SMS Sent                        2745
Olark Chat Conversation          973
Page Visited on Website          640
Converted to Lead                428
Email Bounced                    326
Email Link Clicked               267
Form Submitted on Website        116
NaN                              103
Unreachable                       93
Unsubscribed                      61
Had a Phone Conversation          30
Approached upfront                 9
View in browser link Clicked       6
Email Received                     2
Email Marked Spam                  2
Visited Booth in Tradeshow         1
Resubscribed to emails             1
Name: Last Activity, dtype: int64
```

```
leads['Country'].value_counts(dropna=False)
```

```
India                   6492
NaN                     2461
United States             69
United Arab Emirates      53
Singapore                 24
Saudi Arabia              21
United Kingdom            15
Australia                 13
Qatar                     10
Bahrain                    7
Hong Kong                  7
```

```
Oman                           6
France                         6
unknown                        5
Kuwait                         4
South Africa                   4
Canada                         4
Nigeria                        4
Germany                        4
Sweden                         3
Philippines                    2
Uganda                         2
Italy                          2
Bangladesh                     2
Netherlands                    2
Asia/Pacific Region            2
China                          2
Belgium                        2
Ghana                          2
Kenya                          1
Sri Lanka                      1
Tanzania                       1
Malaysia                       1
Liberia                        1
Switzerland                    1
Denmark                        1
Russia                         1
Vietnam                        1
Indonesia                      1
Name: Country, dtype: int64
```

```
leads['Specialization'].value_counts(dropna=False)
```

```
NaN                                3380
Finance Management                  976
Human Resource Management           848
Marketing Management                838
Operations Management               503
Business Administration             403
IT Projects Management              366
Supply Chain Management             349
Banking, Investment And Insurance   338
Travel and Tourism                  203
Media and Advertising               203
International Business              178
Healthcare Management               159
Hospitality Management              114
E-COMMERCE                          112
Retail Management                   100
Rural and Agribusiness               73
E-Business                           57
```

```
Services Excellence                    40
Name: Specialization, dtype: int64
```

```
leads['How did you hear about X Education'].value_counts(dropna=False)
```

```
NaN                    7250
Online Search           808
Word Of Mouth           348
Student of SomeSchool   310
Other                   186
Multiple Sources        152
Advertisements           70
Social Media             67
Email                    26
SMS                      23
Name: How did you hear about X Education, dtype: int64
```

```
leads['What is your current occupation'].value_counts(dropna=False)
```

```
Unemployed              5600
NaN                     2690
Working Professional     706
Student                  210
Other                     16
Housewife                 10
Businessman                8
Name: What is your current occupation, dtype: int64
```

```
leads['What matters most to you in choosing a
course'].value_counts(dropna=False)
```

```
Better Career Prospects      6528
NaN                          2709
Flexibility & Convenience       2
Other                           1
Name: What matters most to you in choosing a course, dtype: int64
```

```
leads['Search'].value_counts(dropna=False) # 14/9240 yes => drop
```

```
No     9226
Yes      14
Name: Search, dtype: int64
```

```
leads['Magazine'].value_counts(dropna=False) # => all of Magazine is
No, drop this column
```

```
No    9240
Name: Magazine, dtype: int64
```

```
leads['Newspaper Article'].value_counts(dropna=False) # => Newspaper
Article Yes is 2 only, drop this column
```

```
No     9238
Yes       2
Name: Newspaper Article, dtype: int64
```

```
leads['X Education Forums'].value_counts(dropna=False) # => X Education Forums Yes is 1 only, drop this column
```

```
No     9239
Yes       1
Name: X Education Forums, dtype: int64
```

```
leads['Newspaper'].value_counts(dropna=False) # Newspaper 1/9240 => drop
```

```
No     9239
Yes       1
Name: Newspaper, dtype: int64
```

```
leads['Digital Advertisement'].value_counts(dropna=False) # Yes 4/9240 => Drop
```

```
No     9236
Yes       4
Name: Digital Advertisement, dtype: int64
```

```
leads['Through Recommendations'].value_counts(dropna=False) # Yes 7/9240 => Drop
```

```
No     9233
Yes       7
Name: Through Recommendations, dtype: int64
```

```
leads['Receive More Updates About Our Courses'].value_counts(dropna=False) # All no => drop
```

```
No    9240
Name: Receive More Updates About Our Courses, dtype: int64
```

```
leads['Tags'].value_counts(dropna=False)
```

```
NaN                                                  3353
Will revert after reading the email                  2072
Ringing                                              1203
Interested in other courses                           513
Already a student                                     465
Closed by Horizzon                                    358
switched off                                          240
Busy                                                  186
Lost to EINS                                          175
Not doing further education                           145
Interested  in full time MBA                          117
Graduation in progress                                111
```

```
invalid number                                        83
Diploma holder (Not Eligible)                         63
wrong number given                                    47
opp hangup                                            33
number not provided                                   27
in touch with EINS                                    12
Lost to Others                                         7
Still Thinking                                         6
Want to take admission but has financial problems      6
In confusion whether part time or DLP                  5
Interested in Next batch                               5
Lateral student                                        3
Shall take in the next coming month                    2
University not recognized                              2
Recognition issue (DEC approval)                       1
Name: Tags, dtype: int64
```

```
leads['Update me on Supply Chain Content'].value_counts(dropna=False)
# No 9240 => Drop
```

```
No    9240
Name: Update me on Supply Chain Content, dtype: int64
```

```
leads['Get updates on DM Content'].value_counts(dropna=False) # No
9240 => Drop
```

```
No    9240
Name: Get updates on DM Content, dtype: int64
```

```
leads['Lead Profile'].value_counts(dropna=False)
```

```
NaN                              6855
Potential Lead                   1613
Other Leads                       487
Student of SomeSchool             241
Lateral Student                    24
Dual Specialization Student        20
Name: Lead Profile, dtype: int64
```

```
leads['City'].value_counts(dropna=False)
```

```
NaN                              3669
Mumbai                           3222
Thane & Outskirts                 752
Other Cities                      686
Other Cities of Maharashtra       457
Other Metro Cities                380
Tier II Cities                     74
Name: City, dtype: int64
```

```
leads['I agree to pay the amount through
cheque'].value_counts(dropna=False) # No 9240 => Drop
```

```
No    9240
Name: I agree to pay the amount through cheque, dtype: int64
```

```
leads['A free copy of Mastering The
Interview'].value_counts(dropna=False) # Binary var
```

```
No     6352
Yes    2888
Name: A free copy of Mastering The Interview, dtype: int64
```

```
leads['Last Notable Activity'].value_counts(dropna=False)
```

```
Modified                        3407
Email Opened                    2827
SMS Sent                        2172
Page Visited on Website          318
Olark Chat Conversation          183
Email Link Clicked               173
Email Bounced                     60
Unsubscribed                      47
Unreachable                       32
Had a Phone Conversation          14
Email Marked Spam                  2
Approached upfront                 1
Resubscribed to emails             1
View in browser link Clicked       1
Form Submitted on Website          1
Email Received                     1
Name: Last Notable Activity, dtype: int64
```

Drop columns with all No value, or too less Yes

```
no_columns = no_columns = ['Do Not Call', 'Search', 'Magazine',
'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital
Advertisement', 'Through Recommendations', 'Receive More Updates About
Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM
Content', 'I agree to pay the amount through cheque', 'Do Not Email']
leads.drop(no_columns, 1, inplace=True)
leads.head()
```

```
                Lead Origin     Lead Source  Converted  TotalVisits  \
0                       API      Olark Chat          0          0.0
1                       API  Organic Search          0          5.0
2  Landing Page Submission  Direct Traffic          1          2.0
3  Landing Page Submission  Direct Traffic          0          1.0
4  Landing Page Submission          Google          1          2.0

   Total Time Spent on Website  Page Views Per Visit            Last
```

```
Activity  \
0                            0                   0.0  Page Visited on
Website
1                          674                   2.5             Email
Opened
2                         1532                   2.0             Email
Opened
3                          305                   1.0
Unreachable
4                         1428                   1.0        Converted
to Lead

  Country           Specialization How did you hear about X Education
\
0     NaN                      NaN                                  NaN

1   India                      NaN                                  NaN

2   India  Business Administration                                  NaN

3   India    Media and Advertising                     Word Of Mouth

4   India                      NaN                                Other


  What is your current occupation  \
0                      Unemployed
1                      Unemployed
2                         Student
3                      Unemployed
4                      Unemployed

  What matters most to you in choosing a course  \
0                       Better Career Prospects
1                       Better Career Prospects
2                       Better Career Prospects
3                       Better Career Prospects
4                       Better Career Prospects

                                  Tags    Lead Profile    City  \
0          Interested in other courses             NaN     NaN
1                              Ringing             NaN     NaN
2  Will revert after reading the email  Potential Lead  Mumbai
3                              Ringing             NaN  Mumbai
4  Will revert after reading the email             NaN  Mumbai

  A free copy of Mastering The Interview Last Notable Activity
0                                     No              Modified
1                                     No          Email Opened
2                                    Yes          Email Opened
```

```
3                                        No              Modified
4                                        No              Modified

leads.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 17 columns):
 #   Column                                         Non-Null Count
Dtype
---  ------                                         --------------
-----
 0   Lead Origin                                    9240 non-null
object
 1   Lead Source                                    9204 non-null
object
 2   Converted                                      9240 non-null
int64
 3   TotalVisits                                    9103 non-null
float64
 4   Total Time Spent on Website                    9240 non-null
int64
 5   Page Views Per Visit                           9103 non-null
float64
 6   Last Activity                                  9137 non-null
object
 7   Country                                        6779 non-null
object
 8   Specialization                                 5860 non-null
object
 9   How did you hear about X Education             1990 non-null
object
 10  What is your current occupation                6550 non-null
object
 11  What matters most to you in choosing a course  6531 non-null
object
 12  Tags                                           5887 non-null
object
 13  Lead Profile                                   2385 non-null
object
 14  City                                           5571 non-null
object
 15  A free copy of Mastering The Interview         9240 non-null
object
 16  Last Notable Activity                          9240 non-null
object
dtypes: float64(2), int64(2), object(13)
memory usage: 1.2+ MB
```

Handle Columns which have NaN/Select values:

Total rows: 9240

- Lead Source: 36 => impute them with Google
- Last Activity: 103 => impute them with Email Opened
- Country: 2461 => X Edutaion is base in India, impute them with India
- Specialization: Select: 1942, NaN: 1438 => impute them with Others
- How did you hear about X Education: Select: 5043, NaN: 2207 => 7250 NaN/Select over 9240 => drop
- What is your current occupation: 2690 => impute them with Unemployed
- What matters most to you in choosing a course: 2709 => this column seems not a valuable in model => drop
- Tags: 3353 => this column seems not a valuable in model => drop
- Lead Profile: Select: 4146, NaN: 2709 => 6855 Select/NaN values over 9240 => drop
- City: Select: 2249, NaN: 1420 => impute with Mumbai

```
# Drop un-used columns
unused_columns = ['How did you hear about X Education', 'What matters
most to you in choosing a course', 'Tags', 'Lead Profile']
leads.drop(unused_columns, 1, inplace=True)
leads.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 13 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Lead Origin                             9240 non-null   object
 1   Lead Source                             9204 non-null   object
 2   Converted                               9240 non-null   int64
 3   TotalVisits                             9103 non-null   float64
 4   Total Time Spent on Website             9240 non-null   int64
 5   Page Views Per Visit                    9103 non-null   float64
 6   Last Activity                           9137 non-null   object
 7   Country                                 6779 non-null   object
 8   Specialization                          5860 non-null   object
 9   What is your current occupation         6550 non-null   object
 10  City                                    5571 non-null   object
 11  A free copy of Mastering The Interview  9240 non-null   object
 12  Last Notable Activity                   9240 non-null   object
dtypes: float64(2), int64(2), object(9)
memory usage: 938.6+ KB
```

```
# Impute columns
leads['Lead Source']= leads['Lead Source'].replace(np.nan, 'Google')
leads['Last Activity']= leads['Last Activity'].replace(np.nan, 'Email
Opened')
leads['Country']= leads['Country'].replace(np.nan, 'India')
```

```
leads['Specialization']= leads['Specialization'].replace(np.nan,
'Others')
leads['What is your current occupation']= leads['What is your current
occupation'].replace(np.nan, 'Unemployed')
leads['City']= leads['City'].replace(np.nan, 'Mumbai')
leads.describe()
```

```
         Converted  TotalVisits  Total Time Spent on Website  \
count  9240.000000  9103.000000                   9240.000000
mean      0.385390     3.445238                    487.698268
std       0.486714     4.854853                    548.021466
min       0.000000     0.000000                      0.000000
25%       0.000000     1.000000                     12.000000
50%       0.000000     3.000000                    248.000000
75%       1.000000     5.000000                    936.000000
max       1.000000   251.000000                   2272.000000

       Page Views Per Visit
count           9103.000000
mean               2.362820
std                2.161418
min                0.000000
25%                1.000000
50%                2.000000
75%                3.000000
max               55.000000
```

```
leads.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 13 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Lead Origin                             9240 non-null   object
 1   Lead Source                             9240 non-null   object
 2   Converted                               9240 non-null   int64
 3   TotalVisits                             9103 non-null   float64
 4   Total Time Spent on Website             9240 non-null   int64
 5   Page Views Per Visit                    9103 non-null   float64
 6   Last Activity                           9240 non-null   object
 7   Country                                 9240 non-null   object
 8   Specialization                          9240 non-null   object
 9   What is your current occupation         9240 non-null   object
 10  City                                    9240 non-null   object
 11  A free copy of Mastering The Interview  9240 non-null   object
 12  Last Notable Activity                   9240 non-null   object
dtypes: float64(2), int64(2), object(9)
memory usage: 938.6+ KB
```

Map Yes/No as 1/0 for Binary Variables

```
# Defining the map function
def binary_map(x):
    return x.map({'Yes': 1, "No": 0})

# List of Yes/No Binary Variables
binary_columns = ['A free copy of Mastering The Interview']

# Apply map
leads[binary_columns] = leads[binary_columns].apply(binary_map)

leads.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 13 columns):
 #   Column                                  Non-Null Count  Dtype  
---  ------                                  --------------  -----  
 0   Lead Origin                             9240 non-null   object 
 1   Lead Source                             9240 non-null   object 
 2   Converted                               9240 non-null   int64  
 3   TotalVisits                             9103 non-null   float64
 4   Total Time Spent on Website             9240 non-null   int64  
 5   Page Views Per Visit                    9103 non-null   float64
 6   Last Activity                           9240 non-null   object 
 7   Country                                 9240 non-null   object 
 8   Specialization                          9240 non-null   object 
 9   What is your current occupation         9240 non-null   object 
 10  City                                    9240 non-null   object 
 11  A free copy of Mastering The Interview  9240 non-null   int64  
 12  Last Notable Activity                   9240 non-null   object 
dtypes: float64(2), int64(3), object(8)
memory usage: 938.6+ KB
```

Rename columns for better view

```
rename_columns = {'Lead Origin': 'LeadOrigin', 'Lead Source':
'LeadSource', 'Total Time Spent on Website':
'TotalTimeSpentOnWebsite', 'Page Views Per Visit':
'PageViewsPerVisit',
                  'Last Activity': 'LastActivity', 'What is your
current occupation': 'CurrentOccupation', 'A free copy of Mastering
The Interview': 'FreeCopyOfMasteringInterview',
                  'Last Notable Activity': 'LastNotableActivity'}
leads.rename(columns=rename_columns, inplace=True)
leads.head()
```

```
                LeadOrigin      LeadSource  Converted  TotalVisits  \
0                      API      Olark Chat          0          0.0   
1                      API  Organic Search          0          5.0   
```

```
2  Landing Page Submission  Direct Traffic          1          2.0
3  Landing Page Submission  Direct Traffic          0          1.0
4  Landing Page Submission          Google          1          2.0

   TotalTimeSpentOnWebsite  PageViewsPerVisit             LastActivity
\
0                        0                0.0  Page Visited on Website

1                      674                2.5             Email Opened

2                     1532                2.0             Email Opened

3                      305                1.0              Unreachable

4                     1428                1.0        Converted to Lead


  Country           Specialization CurrentOccupation    City  \
0   India                   Others        Unemployed  Mumbai
1   India                   Others        Unemployed  Mumbai
2   India  Business Administration           Student  Mumbai
3   India    Media and Advertising        Unemployed  Mumbai
4   India                   Others        Unemployed  Mumbai

   FreeCopyOfMasteringInterview LastNotableActivity
0                             0            Modified
1                             0        Email Opened
2                             1        Email Opened
3                             0            Modified
4                             0            Modified
```

Handle low frequency values

```
#replacing Nan Values and combining low frequency values
leads.LeadSource = leads.LeadSource.replace(np.nan,'Others')
leads.LeadSource = leads.LeadSource.replace('google','Google')
leads.LeadSource = leads.LeadSource.replace('Facebook','Social Media')
leads.LeadSource =
leads.LeadSource.replace(['bing','Click2call','Press_Release',

'youtubechannel','welearnblog_Home',

'WeLearn','blog','Pay per Click Ads',

'testone','NC_EDM', 'Live Chat'] ,'Others')

leads.LeadSource.value_counts()
```

```
Google              2909
Direct Traffic      2543
Olark Chat          1755
```

```
Organic Search      1154
Reference            534
Welingak Website     142
Referral Sites       125
Social Media          57
Others                21
Name: LeadSource, dtype: int64
```

```
#replacing Nan Values and combining low frequency values

leads.LastActivity = leads.LastActivity.replace(np.nan,'Others')
leads.LastActivity =
leads.LastActivity.replace(['Unreachable','Unsubscribed',
                                                        'Had a Phone
Conversation',
                                                        'Approached
upfront',
                                                        'View in
browser link Clicked',
                                                        'Email Marked
Spam',
                                                        'Email
Received','Resubscribed to emails',
                                                         'Visited
Booth in Tradeshow'],'Others')

leads.LastActivity.value_counts()
```

```
Email Opened                 3399
SMS Sent                     2709
Olark Chat Conversation       966
Page Visited on Website       597
Converted to Lead             428
Email Bounced                 310
Email Link Clicked            265
Others                        188
Form Submitted on Website     114
Name: LastActivity, dtype: int64
```

Check Outliers for Numeric Variables

```
leads.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 13 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   LeadOrigin                    9240 non-null   object
```

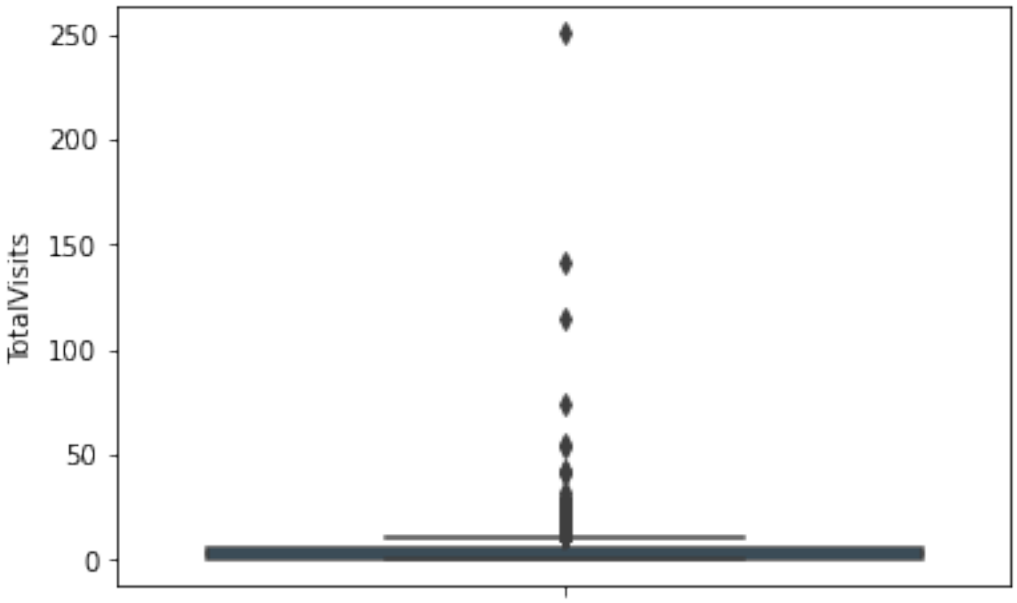
```
 1   LeadSource                     9240 non-null   object
 2   Converted                      9240 non-null   int64
 3   TotalVisits                    9103 non-null   float64
 4   TotalTimeSpentOnWebsite        9240 non-null   int64
 5   PageViewsPerVisit              9103 non-null   float64
 6   LastActivity                   9240 non-null   object
 7   Country                        9240 non-null   object
 8   Specialization                 9240 non-null   object
 9   CurrentOccupation              9240 non-null   object
 10  City                           9240 non-null   object
 11  FreeCopyOfMasteringInterview   9240 non-null   int64
 12  LastNotableActivity            9240 non-null   object
dtypes: float64(2), int64(3), object(8)
memory usage: 938.6+ KB
```

```
numeric_columns = ['TotalVisits', 'TotalTimeSpentOnWebsite',
'PageViewsPerVisit']
leads[numeric_columns].describe(percentiles=[.25, .5, .75, .90, .95, .
99])
```

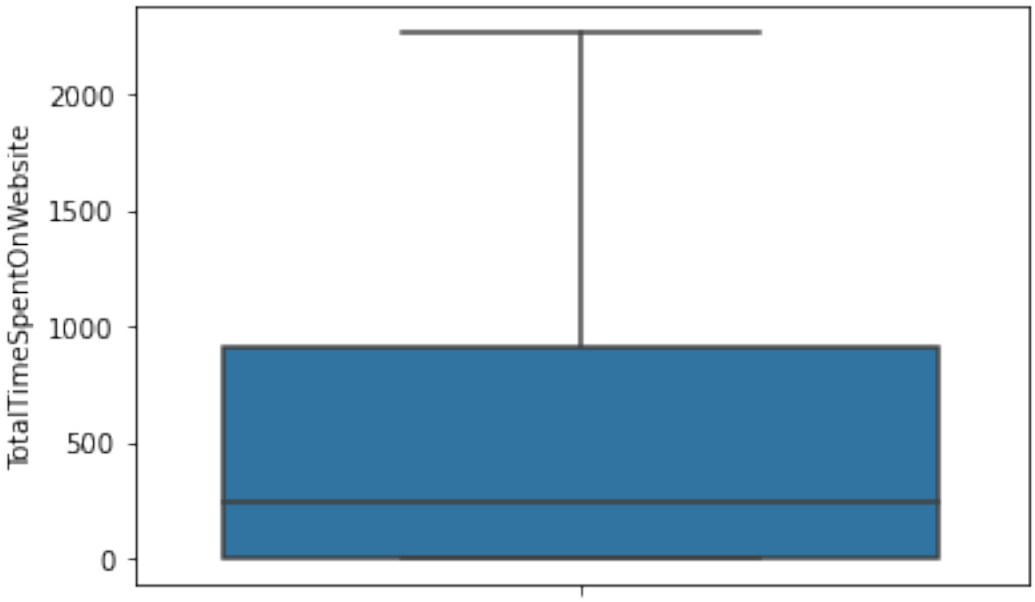
```
       TotalVisits  TotalTimeSpentOnWebsite  PageViewsPerVisit
count  9103.000000              9240.000000        9103.000000
mean      3.445238               487.698268           2.362820
std       4.854853               548.021466           2.161418
min       0.000000                 0.000000           0.000000
25%       1.000000                12.000000           1.000000
50%       3.000000               248.000000           2.000000
75%       5.000000               936.000000           3.000000
90%       7.000000              1380.000000           5.000000
95%      10.000000              1562.000000           6.000000
99%      17.000000              1840.610000           9.000000
max     251.000000              2272.000000          55.000000
```

Deal with outliers

```
sns.boxplot(y=leads['TotalVisits'])
plt.show()
```



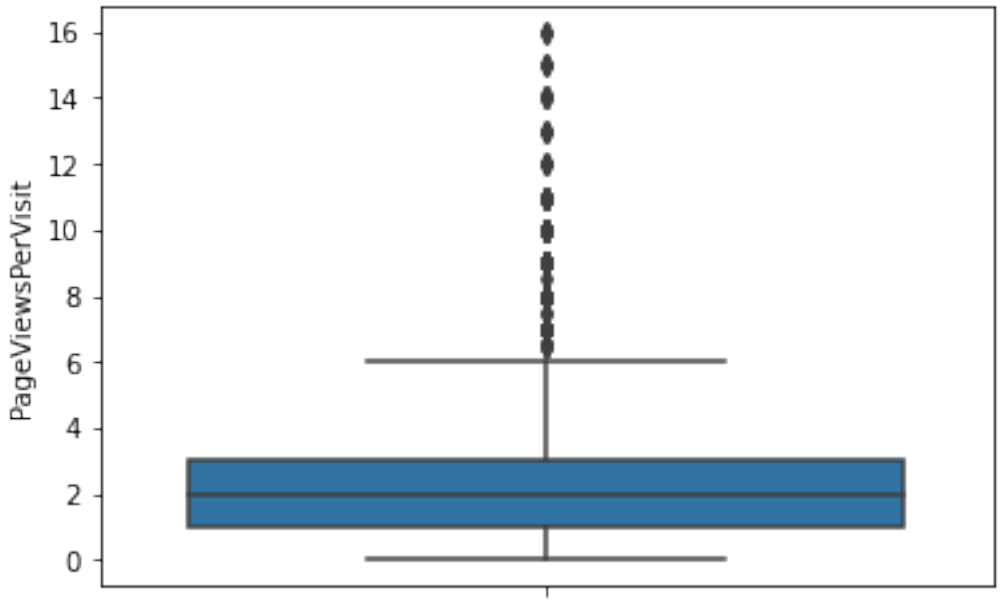

```
leads = leads[(leads.TotalVisits <= leads.TotalVisits.quantile(.99))]

sns.boxplot(y=leads['TotalTimeSpentOnWebsite'])
plt.show()
```




```
leads[(leads.TotalTimeSpentOnWebsite <= 1500)].shape
```

```
(8456, 13)
```

```
sns.boxplot(y=leads.PageViewsPerVisit)
plt.show()
```




```
leads.PageViewsPerVisit.describe()
```

```
count    9020.000000
mean        2.337271
std         2.062363
min         0.000000
25%         1.000000
50%         2.000000
75%         3.000000
max        16.000000
Name: PageViewsPerVisit, dtype: float64
```

```
leads = leads[(leads.PageViewsPerVisit <= 10)]
```

```
leads.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8976 entries, 0 to 9239
Data columns (total 13 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   LeadOrigin                      8976 non-null   object 
 1   LeadSource                      8976 non-null   object 
 2   Converted                       8976 non-null   int64  
 3   TotalVisits                     8976 non-null   float64
 4   TotalTimeSpentOnWebsite         8976 non-null   int64  
 5   PageViewsPerVisit               8976 non-null   float64
 6   LastActivity                    8976 non-null   object 
 7   Country                         8976 non-null   object 
 8   Specialization                  8976 non-null   object 
 9   CurrentOccupation               8976 non-null   object 
```

```
 10  City                           8976 non-null   object
 11  FreeCopyOfMasteringInterview  8976 non-null   int64
 12  LastNotableActivity           8976 non-null   object
dtypes: float64(2), int64(3), object(8)
memory usage: 981.8+ KB
```

Deal with Categorical Variables

```
leads.head()
```

```
                LeadOrigin      LeadSource  Converted  TotalVisits  \
0                      API      Olark Chat          0          0.0
1                      API  Organic Search          0          5.0
2  Landing Page Submission  Direct Traffic          1          2.0
3  Landing Page Submission  Direct Traffic          0          1.0
4  Landing Page Submission          Google          1          2.0

   TotalTimeSpentOnWebsite  PageViewsPerVisit             LastActivity
\
0                        0                0.0  Page Visited on Website

1                      674                2.5             Email Opened

2                     1532                2.0             Email Opened

3                      305                1.0              Unreachable

4                     1428                1.0        Converted to Lead


  Country           Specialization CurrentOccupation    City  \
0   India                   Others        Unemployed  Mumbai
1   India                   Others        Unemployed  Mumbai
2   India  Business Administration           Student  Mumbai
3   India    Media and Advertising        Unemployed  Mumbai
4   India                   Others        Unemployed  Mumbai

   FreeCopyOfMasteringInterview LastNotableActivity
0                             0            Modified
1                             0        Email Opened
2                             1        Email Opened
3                             0            Modified
4                             0            Modified
```

```
leads.LeadSource.value_counts()
```

```
Google              2875
Direct Traffic      2505
Olark Chat          1751
Organic Search      1103
Reference            442
```

```
Welingak Website     129
Referral Sites       118
Social Media          33
Others                20
Name: LeadSource, dtype: int64
```

```
# city & country are correrlated, let's drop city
leads.drop(['City'], 1, inplace=True)
leads.head()
```

	LeadOrigin	LeadSource	Converted	TotalVisits	TotalTimeSpentOnWebsite	PageViewsPerVisit	LastActivity
0	API	Olark Chat	0	0.0	0	0.0	Page Visited on Website
1	API	Organic Search	0	5.0	674	2.5	Email Opened
2	Landing Page Submission	Direct Traffic	1	2.0	1532	2.0	Email Opened
3	Landing Page Submission	Direct Traffic	0	1.0	305	1.0	Unreachable
4	Landing Page Submission	Google	1	2.0	1428	1.0	Converted to Lead

	Country	Specialization	CurrentOccupation	FreeCopyOfMasteringInterview	LastNotableActivity
0	India	Others	Unemployed	0	Modified
1	India	Others	Unemployed	0	Email Opened
2	India	Business Administration	Student	1	Email Opened
3	India	Media and Advertising	Unemployed	0	Modified
4	India	Others	Unemployed	0	Modified

```
# Generate dummy variables
dummy = pd.get_dummies(leads[['LeadOrigin','CurrentOccupation']],
drop_first=True)
leads = pd.concat([leads,dummy],1)
leads.head()
```

```
                LeadOrigin      LeadSource  Converted  TotalVisits  \
0                      API      Olark Chat          0          0.0
1                      API  Organic Search          0          5.0
2  Landing Page Submission  Direct Traffic          1          2.0
3  Landing Page Submission  Direct Traffic          0          1.0
4  Landing Page Submission          Google          1          2.0

   TotalTimeSpentOnWebsite  PageViewsPerVisit             LastActivity
\
0                        0                0.0  Page Visited on Website

1                      674                2.5             Email Opened

2                     1532                2.0             Email Opened

3                      305                1.0              Unreachable

4                     1428                1.0        Converted to Lead


  Country           Specialization CurrentOccupation  \
0   India                   Others        Unemployed
1   India                   Others        Unemployed
2   India  Business Administration           Student
3   India    Media and Advertising        Unemployed
4   India                   Others        Unemployed

   FreeCopyOfMasteringInterview LastNotableActivity  \
0                             0            Modified
1                             0        Email Opened
2                             1        Email Opened
3                             0            Modified
4                             0            Modified

   LeadOrigin_Landing Page Submission  LeadOrigin_Lead Add Form  \
0                                   0                         0
1                                   0                         0
2                                   1                         0
3                                   1                         0
4                                   1                         0

   LeadOrigin_Lead Import  CurrentOccupation_Housewife  \
0                       0                            0
1                       0                            0
2                       0                            0
3                       0                            0
4                       0                            0

   CurrentOccupation_Other  CurrentOccupation_Student  \
0                        0                          0
1                        0                          0
```

```
2                           0                           1
3                           0                           0
4                           0                           0

   CurrentOccupation_Unemployed  CurrentOccupation_Working
Professional
0                             1
0
1                             1
0
2                             0
0
3                             1
0
4                             1
0

dummy = pd.get_dummies(leads['Specialization'],
prefix='Specialization')
dummy = dummy.drop(['Specialization_Others'], 1)
leads = pd.concat([leads, dummy], axis = 1)
leads.head()

                LeadOrigin      LeadSource  Converted  TotalVisits  \
0                      API      Olark Chat          0          0.0
1                      API  Organic Search          0          5.0
2  Landing Page Submission  Direct Traffic          1          2.0
3  Landing Page Submission  Direct Traffic          0          1.0
4  Landing Page Submission          Google          1          2.0

   TotalTimeSpentOnWebsite  PageViewsPerVisit             LastActivity
\
0                        0                0.0  Page Visited on Website

1                      674                2.5             Email Opened

2                     1532                2.0             Email Opened

3                      305                1.0              Unreachable

4                     1428                1.0        Converted to Lead


  Country           Specialization CurrentOccupation  ...  \
0   India                   Others        Unemployed  ...
1   India                   Others        Unemployed  ...
2   India  Business Administration           Student  ...
3   India    Media and Advertising        Unemployed  ...
4   India                   Others        Unemployed  ...

   Specialization_IT Projects Management  \
```

```
0                                          0
1                                          0
2                                          0
3                                          0
4                                          0

  Specialization_International Business  Specialization_Marketing
Management  \
0                                        0
0
1                                        0
0
2                                        0
0
3                                        0
0
4                                        0
0

   Specialization_Media and Advertising  Specialization_Operations
Management  \
0                                        0
0
1                                        0
0
2                                        0
0
3                                        1
0
4                                        0
0

   Specialization_Retail Management  Specialization_Rural and
Agribusiness  \
0                                   0
0
1                                   0
0
2                                   0
0
3                                   0
0
4                                   0
0

   Specialization_Services Excellence  Specialization_Supply Chain
Management  \
0                                     0
0
1                                     0
```

```
0
2                                    0
0
3                                    0
0
4                                    0
0

   Specialization_Travel and Tourism
0                                    0
1                                    0
2                                    0
3                                    0
4                                    0

[5 rows x 38 columns]
```

```
dummy = pd.get_dummies(leads.LeadSource, prefix='LeadSource')
dummy = dummy.drop(['LeadSource_Others'], 1)
leads = pd.concat([leads, dummy], axis = 1)
leads.head()
```

```
                LeadOrigin      LeadSource  Converted  TotalVisits  \
0                      API      Olark Chat          0          0.0
1                      API  Organic Search          0          5.0
2  Landing Page Submission  Direct Traffic          1          2.0
3  Landing Page Submission  Direct Traffic          0          1.0
4  Landing Page Submission          Google          1          2.0

   TotalTimeSpentOnWebsite  PageViewsPerVisit             LastActivity
\
0                        0                0.0  Page Visited on Website

1                      674                2.5             Email Opened

2                     1532                2.0             Email Opened

3                      305                1.0              Unreachable

4                     1428                1.0        Converted to Lead


  Country           Specialization CurrentOccupation  ...  \
0   India                   Others        Unemployed  ...
1   India                   Others        Unemployed  ...
2   India  Business Administration           Student  ...
3   India    Media and Advertising        Unemployed  ...
4   India                   Others        Unemployed  ...

   Specialization_Supply Chain Management Specialization_Travel and
Tourism  \
```

```
0                                          0
0
1                                          0
0
2                                          0
0
3                                          0
0
4                                          0
0

   LeadSource_Direct Traffic  LeadSource_Google  LeadSource_Olark Chat
\
0                          0                 0                      1

1                          0                 0                      0

2                          1                 0                      0

3                          1                 0                      0

4                          0                 1                      0


   LeadSource_Organic Search  LeadSource_Reference
LeadSource_Referral Sites  \
0                          0                     0
0
1                          1                     0
0
2                          0                     0
0
3                          0                     0
0
4                          0                     0
0

   LeadSource_Social Media  LeadSource_Welingak Website
0                        0                            0
1                        0                            0
2                        0                            0
3                        0                            0
4                        0                            0

[5 rows x 46 columns]
```

```
dummy = pd.get_dummies(leads.LastActivity, prefix='LastActivity')
dummy = dummy.drop(['LastActivity_Others'], 1)
leads = pd.concat([leads, dummy], axis = 1)
leads.head()
```

```
                 LeadOrigin      LeadSource  Converted  TotalVisits  \
0                       API      Olark Chat          0          0.0
1                       API  Organic Search          0          5.0
2  Landing Page Submission  Direct Traffic          1          2.0
3  Landing Page Submission  Direct Traffic          0          1.0
4  Landing Page Submission          Google          1          2.0

   TotalTimeSpentOnWebsite  PageViewsPerVisit             LastActivity
\
0                        0                0.0  Page Visited on Website

1                      674                2.5             Email Opened

2                     1532                2.0             Email Opened

3                      305                1.0                   Others

4                     1428                1.0        Converted to Lead


  Country           Specialization CurrentOccupation  ...  \
0   India                   Others        Unemployed  ...
1   India                   Others        Unemployed  ...
2   India  Business Administration           Student  ...
3   India    Media and Advertising        Unemployed  ...
4   India                   Others        Unemployed  ...

   LeadSource_Social Media LeadSource_Welingak Website  \
0                        0                            0
1                        0                            0
2                        0                            0
3                        0                            0
4                        0                            0

   LastActivity_Converted to Lead  LastActivity_Email Bounced  \
0                               0                           0
1                               0                           0
2                               0                           0
3                               0                           0
4                               1                           0

   LastActivity_Email Link Clicked  LastActivity_Email Opened  \
0                                0                          0
1                                0                          1
2                                0                          1
3                                0                          0
4                                0                          0

   LastActivity_Form Submitted on Website  \
0                                       0
1                                       0
```

```
2                                           0
3                                           0
4                                           0

   LastActivity_Olark Chat Conversation  LastActivity_Page Visited on
Website  \
0                                       0
1
1                                       0
0
2                                       0
0
3                                       0
0
4                                       0
0

   LastActivity_SMS Sent
0                      0
1                      0
2                      0
3                      0
4                      0

[5 rows x 54 columns]
```

```
dummy = pd.get_dummies(leads.Country, prefix='Country')
dummy = dummy.drop(['Country_Indonesia'], 1)
leads = pd.concat([leads, dummy], axis = 1)
leads.head()
```

```
   Converted  TotalVisits  TotalTimeSpentOnWebsite  PageViewsPerVisit
Country  \
0          0          0.0                        0                0.0
India
1          0          5.0                      674                2.5
India
2          1          2.0                     1532                2.0
India
3          0          1.0                      305                1.0
India
4          1          2.0                     1428                1.0
India

   FreeCopyOfMasteringInterview LastNotableActivity  \
0                             0            Modified
1                             0       Email Opened
2                             1       Email Opened
3                             0            Modified
4                             0            Modified
```

```
   LeadOrigin_Landing Page Submission  LeadOrigin_Lead Add Form  \
0                                   0                         0
1                                   0                         0
2                                   1                         0
3                                   1                         0
4                                   1                         0

   LeadOrigin_Lead Import  ...  Country_Sri Lanka  Country_Sweden  \
0                       0  ...                  0               0
1                       0  ...                  0               0
2                       0  ...                  0               0
3                       0  ...                  0               0
4                       0  ...                  0               0

   Country_Switzerland  Country_Tanzania  Country_Uganda  \
0                    0                 0               0
1                    0                 0               0
2                    0                 0               0
3                    0                 0               0
4                    0                 0               0

   Country_United Arab Emirates  Country_United Kingdom  \
0                             0                       0
1                             0                       0
2                             0                       0
3                             0                       0
4                             0                       0

   Country_United States  Country_Vietnam  Country_unknown
0                      0                0                0
1                      0                0                0
2                      0                0                0
3                      0                0                0
4                      0                0                0

[5 rows x 86 columns]
```

```
dummy = pd.get_dummies(leads.LastNotableActivity,
prefix='LastNotableActivity')
dummy = dummy.drop(['LastNotableActivity_Email Received'], 1)
leads = pd.concat([leads, dummy], axis = 1)
leads.head()
```

```
   Converted  TotalVisits  TotalTimeSpentOnWebsite  PageViewsPerVisit
Country  \
0          0          0.0                        0                0.0
India
1          0          5.0                      674                2.5
India
```

```
2          1          2.0                      1532                  2.0
India
3          0          1.0                       305                  1.0
India
4          1          2.0                      1428                  1.0
India

   FreeCopyOfMasteringInterview LastNotableActivity  \
0                             0            Modified
1                             0        Email Opened
2                             1        Email Opened
3                             0            Modified
4                             0            Modified

   LeadOrigin_Landing Page Submission  LeadOrigin_Lead Add Form  \
0                                   0                         0
1                                   0                         0
2                                   1                         0
3                                   1                         0
4                                   1                         0

   LeadOrigin_Lead Import  ...  LastNotableActivity_Form Submitted on
Website  \
0                       0  ...
0
1                       0  ...
0
2                       0  ...
0
3                       0  ...
0
4                       0  ...
0

   LastNotableActivity_Had a Phone Conversation
LastNotableActivity_Modified  \
0                                             0
1
1                                             0
0
2                                             0
0
3                                             0
1
4                                             0
1

   LastNotableActivity_Olark Chat Conversation  \
0                                            0
1                                            0
```

```
2                                                    0
3                                                    0
4                                                    0

   LastNotableActivity_Page Visited on Website  \
0                                                    0
1                                                    0
2                                                    0
3                                                    0
4                                                    0

   LastNotableActivity_Resubscribed to emails  LastNotableActivity_SMS
Sent  \
0                                                  0
0
1                                                  0
0
2                                                  0
0
3                                                  0
0
4                                                  0
0

   LastNotableActivity_Unreachable
LastNotableActivity_Unsubscribed  \
0                                  0                                    0

1                                  0                                    0

2                                  0                                    0

3                                  0                                    0

4                                  0                                    0


   LastNotableActivity_View in browser link Clicked
0                                                    0
1                                                    0
2                                                    0
3                                                    0
4                                                    0

[5 rows x 101 columns]

leads.LastNotableActivity.value_counts()

Modified                        3232
Email Opened                    2795
SMS Sent                        2150
```

```
Page Visited on Website          289
Olark Chat Conversation          182
Email Link Clicked               171
Email Bounced                     59
Unsubscribed                      46
Unreachable                       32
Had a Phone Conversation          13
Email Marked Spam                  2
Approached upfront                 1
Resubscribed to emails             1
View in browser link Clicked       1
Form Submitted on Website          1
Email Received                     1
Name: LastNotableActivity, dtype: int64
```

```
# drop original categorical columns
cate_columns = ['LeadOrigin','CurrentOccupation', 'Specialization',
'LeadSource', 'LastActivity', 'Country', 'LastNotableActivity']
leads.drop(cate_columns, 1, inplace=True)
leads.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8976 entries, 0 to 9239
Data columns (total 99 columns):
 #   Column                                        Non-Null Count
Dtype
---  ------                                        --------------
-----
 0   Converted                                     8976 non-null
int64
 1   TotalVisits                                   8976 non-null
float64
 2   TotalTimeSpentOnWebsite                       8976 non-null
int64
 3   PageViewsPerVisit                             8976 non-null
float64
 4   FreeCopyOfMasteringInterview                  8976 non-null
int64
 5   LeadOrigin_Landing Page Submission            8976 non-null
uint8
 6   LeadOrigin_Lead Add Form                      8976 non-null
uint8
 7   LeadOrigin_Lead Import                        8976 non-null
uint8
 8   CurrentOccupation_Housewife                   8976 non-null
uint8
 9   CurrentOccupation_Other                       8976 non-null
uint8
 10  CurrentOccupation_Student                     8976 non-null
uint8
```

```
 11  CurrentOccupation_Unemployed                          8976 non-null
uint8
 12  CurrentOccupation_Working Professional                8976 non-null
uint8
 13  Specialization_Banking, Investment And Insurance      8976 non-null
uint8
 14  Specialization_Business Administration                8976 non-null
uint8
 15  Specialization_E-Business                             8976 non-null
uint8
 16  Specialization_E-COMMERCE                             8976 non-null
uint8
 17  Specialization_Finance Management                     8976 non-null
uint8
 18  Specialization_Healthcare Management                  8976 non-null
uint8
 19  Specialization_Hospitality Management                 8976 non-null
uint8
 20  Specialization_Human Resource Management              8976 non-null
uint8
 21  Specialization_IT Projects Management                 8976 non-null
uint8
 22  Specialization_International Business                 8976 non-null
uint8
 23  Specialization_Marketing Management                   8976 non-null
uint8
 24  Specialization_Media and Advertising                  8976 non-null
uint8
 25  Specialization_Operations Management                  8976 non-null
uint8
 26  Specialization_Retail Management                      8976 non-null
uint8
 27  Specialization_Rural and Agribusiness                 8976 non-null
uint8
 28  Specialization_Services Excellence                    8976 non-null
uint8
 29  Specialization_Supply Chain Management                8976 non-null
uint8
 30  Specialization_Travel and Tourism                     8976 non-null
uint8
 31  LeadSource_Direct Traffic                             8976 non-null
uint8
 32  LeadSource_Google                                     8976 non-null
uint8
 33  LeadSource_Olark Chat                                 8976 non-null
uint8
 34  LeadSource_Organic Search                             8976 non-null
uint8
 35  LeadSource_Reference                                  8976 non-null
```

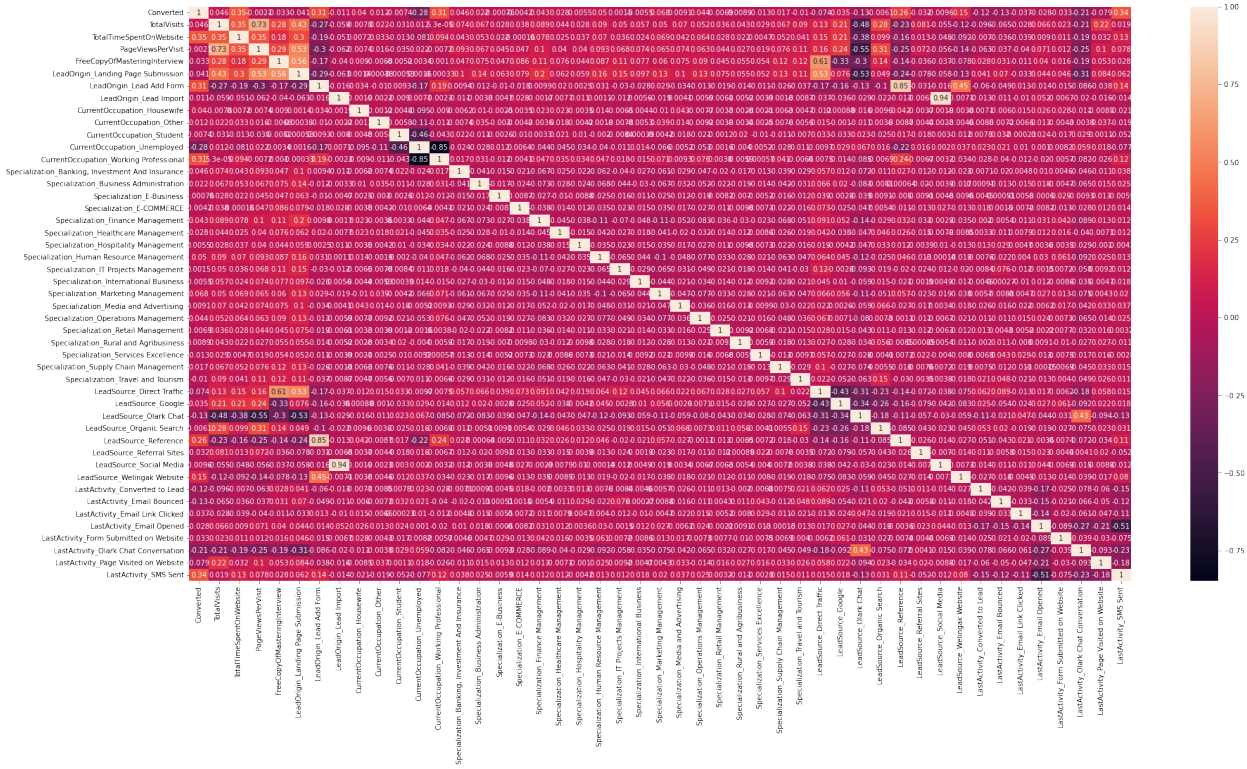
```
uint8
 36  LeadSource_Referral Sites                       8976 non-null
uint8
 37  LeadSource_Social Media                         8976 non-null
uint8
 38  LeadSource_Welingak Website                     8976 non-null
uint8
 39  LastActivity_Converted to Lead                  8976 non-null
uint8
 40  LastActivity_Email Bounced                      8976 non-null
uint8
 41  LastActivity_Email Link Clicked                 8976 non-null
uint8
 42  LastActivity_Email Opened                       8976 non-null
uint8
 43  LastActivity_Form Submitted on Website          8976 non-null
uint8
 44  LastActivity_Olark Chat Conversation            8976 non-null
uint8
 45  LastActivity_Page Visited on Website            8976 non-null
uint8
 46  LastActivity_SMS Sent                           8976 non-null
uint8
 47  Country_Asia/Pacific Region                     8976 non-null
uint8
 48  Country_Australia                               8976 non-null
uint8
 49  Country_Bahrain                                 8976 non-null
uint8
 50  Country_Bangladesh                              8976 non-null
uint8
 51  Country_Belgium                                 8976 non-null
uint8
 52  Country_Canada                                  8976 non-null
uint8
 53  Country_China                                   8976 non-null
uint8
 54  Country_Denmark                                 8976 non-null
uint8
 55  Country_France                                  8976 non-null
uint8
 56  Country_Germany                                 8976 non-null
uint8
 57  Country_Ghana                                   8976 non-null
uint8
 58  Country_Hong Kong                               8976 non-null
uint8
 59  Country_India                                   8976 non-null
uint8
```

```
 60  Country_Italy                                 8976 non-null
uint8
 61  Country_Kenya                                 8976 non-null
uint8
 62  Country_Kuwait                                8976 non-null
uint8
 63  Country_Liberia                               8976 non-null
uint8
 64  Country_Malaysia                              8976 non-null
uint8
 65  Country_Netherlands                           8976 non-null
uint8
 66  Country_Nigeria                               8976 non-null
uint8
 67  Country_Oman                                  8976 non-null
uint8
 68  Country_Philippines                           8976 non-null
uint8
 69  Country_Qatar                                 8976 non-null
uint8
 70  Country_Russia                                8976 non-null
uint8
 71  Country_Saudi Arabia                          8976 non-null
uint8
 72  Country_Singapore                             8976 non-null
uint8
 73  Country_South Africa                          8976 non-null
uint8
 74  Country_Sri Lanka                             8976 non-null
uint8
 75  Country_Sweden                                8976 non-null
uint8
 76  Country_Switzerland                           8976 non-null
uint8
 77  Country_Tanzania                              8976 non-null
uint8
 78  Country_Uganda                                8976 non-null
uint8
 79  Country_United Arab Emirates                  8976 non-null
uint8
 80  Country_United Kingdom                        8976 non-null
uint8
 81  Country_United States                         8976 non-null
uint8
 82  Country_Vietnam                               8976 non-null
uint8
 83  Country_unknown                               8976 non-null
uint8
 84  LastNotableActivity_Approached upfront        8976 non-null
```

```
uint8
 85  LastNotableActivity_Email Bounced                 8976 non-null
uint8
 86  LastNotableActivity_Email Link Clicked            8976 non-null
uint8
 87  LastNotableActivity_Email Marked Spam             8976 non-null
uint8
 88  LastNotableActivity_Email Opened                  8976 non-null
uint8
 89  LastNotableActivity_Form Submitted on Website     8976 non-null
uint8
 90  LastNotableActivity_Had a Phone Conversation      8976 non-null
uint8
 91  LastNotableActivity_Modified                      8976 non-null
uint8
 92  LastNotableActivity_Olark Chat Conversation       8976 non-null
uint8
 93  LastNotableActivity_Page Visited on Website       8976 non-null
uint8
 94  LastNotableActivity_Resubscribed to emails        8976 non-null
uint8
 95  LastNotableActivity_SMS Sent                      8976 non-null
uint8
 96  LastNotableActivity_Unreachable                   8976 non-null
uint8
 97  LastNotableActivity_Unsubscribed                  8976 non-null
uint8
 98  LastNotableActivity_View in browser link Clicked  8976 non-null
uint8
dtypes: float64(2), int64(3), uint8(94)
memory usage: 1.2 MB
```

Looking at Correlations

```
# Let's see the correlation matrix
plt.figure(figsize = (30,15))        # Size of the figure
sns.heatmap(leads.corr(),annot = True)
plt.show()
```



Build Model

Train-Test Split & Logistic Regression Model Building

```
# Putting response variable to y
y = leads['Converted']
y.head()
X=leads.drop('Converted', axis=1)
```

Splitting the data into train and test

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
train_size=0.7, test_size=0.3, random_state=100)

X_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6283 entries, 5555 to 5812
Data columns (total 98 columns):
 #   Column                                                Non-Null Count
Dtype
---  ------                                                --------------
-----
 0   TotalVisits                                           6283 non-null
float64
 1   TotalTimeSpentOnWebsite                               6283 non-null
int64
```

```
 2   PageViewsPerVisit                                 6283 non-null
float64
 3   FreeCopyOfMasteringInterview                      6283 non-null
int64
 4   LeadOrigin_Landing Page Submission                6283 non-null
uint8
 5   LeadOrigin_Lead Add Form                          6283 non-null
uint8
 6   LeadOrigin_Lead Import                            6283 non-null
uint8
 7   CurrentOccupation_Housewife                       6283 non-null
uint8
 8   CurrentOccupation_Other                           6283 non-null
uint8
 9   CurrentOccupation_Student                         6283 non-null
uint8
 10  CurrentOccupation_Unemployed                      6283 non-null
uint8
 11  CurrentOccupation_Working Professional            6283 non-null
uint8
 12  Specialization_Banking, Investment And Insurance  6283 non-null
uint8
 13  Specialization_Business Administration            6283 non-null
uint8
 14  Specialization_E-Business                         6283 non-null
uint8
 15  Specialization_E-COMMERCE                         6283 non-null
uint8
 16  Specialization_Finance Management                 6283 non-null
uint8
 17  Specialization_Healthcare Management              6283 non-null
uint8
 18  Specialization_Hospitality Management             6283 non-null
uint8
 19  Specialization_Human Resource Management          6283 non-null
uint8
 20  Specialization_IT Projects Management             6283 non-null
uint8
 21  Specialization_International Business             6283 non-null
uint8
 22  Specialization_Marketing Management               6283 non-null
uint8
 23  Specialization_Media and Advertising              6283 non-null
uint8
 24  Specialization_Operations Management              6283 non-null
uint8
 25  Specialization_Retail Management                  6283 non-null
uint8
 26  Specialization_Rural and Agribusiness             6283 non-null
```

```
uint8
 27  Specialization_Services Excellence                6283 non-null
uint8
 28  Specialization_Supply Chain Management            6283 non-null
uint8
 29  Specialization_Travel and Tourism                 6283 non-null
uint8
 30  LeadSource_Direct Traffic                         6283 non-null
uint8
 31  LeadSource_Google                                 6283 non-null
uint8
 32  LeadSource_Olark Chat                             6283 non-null
uint8
 33  LeadSource_Organic Search                         6283 non-null
uint8
 34  LeadSource_Reference                              6283 non-null
uint8
 35  LeadSource_Referral Sites                         6283 non-null
uint8
 36  LeadSource_Social Media                           6283 non-null
uint8
 37  LeadSource_Welingak Website                       6283 non-null
uint8
 38  LastActivity_Converted to Lead                    6283 non-null
uint8
 39  LastActivity_Email Bounced                        6283 non-null
uint8
 40  LastActivity_Email Link Clicked                   6283 non-null
uint8
 41  LastActivity_Email Opened                         6283 non-null
uint8
 42  LastActivity_Form Submitted on Website            6283 non-null
uint8
 43  LastActivity_Olark Chat Conversation              6283 non-null
uint8
 44  LastActivity_Page Visited on Website              6283 non-null
uint8
 45  LastActivity_SMS Sent                             6283 non-null
uint8
 46  Country_Asia/Pacific Region                       6283 non-null
uint8
 47  Country_Australia                                 6283 non-null
uint8
 48  Country_Bahrain                                   6283 non-null
uint8
 49  Country_Bangladesh                                6283 non-null
uint8
 50  Country_Belgium                                   6283 non-null
uint8
```

```
 51  Country_Canada                                6283 non-null
uint8
 52  Country_China                                 6283 non-null
uint8
 53  Country_Denmark                               6283 non-null
uint8
 54  Country_France                                6283 non-null
uint8
 55  Country_Germany                               6283 non-null
uint8
 56  Country_Ghana                                 6283 non-null
uint8
 57  Country_Hong Kong                             6283 non-null
uint8
 58  Country_India                                 6283 non-null
uint8
 59  Country_Italy                                 6283 non-null
uint8
 60  Country_Kenya                                 6283 non-null
uint8
 61  Country_Kuwait                                6283 non-null
uint8
 62  Country_Liberia                               6283 non-null
uint8
 63  Country_Malaysia                              6283 non-null
uint8
 64  Country_Netherlands                           6283 non-null
uint8
 65  Country_Nigeria                               6283 non-null
uint8
 66  Country_Oman                                  6283 non-null
uint8
 67  Country_Philippines                           6283 non-null
uint8
 68  Country_Qatar                                 6283 non-null
uint8
 69  Country_Russia                                6283 non-null
uint8
 70  Country_Saudi Arabia                          6283 non-null
uint8
 71  Country_Singapore                             6283 non-null
uint8
 72  Country_South Africa                          6283 non-null
uint8
 73  Country_Sri Lanka                             6283 non-null
uint8
 74  Country_Sweden                                6283 non-null
uint8
 75  Country_Switzerland                           6283 non-null
```

```
uint8
 76  Country_Tanzania                                    6283 non-null
uint8
 77  Country_Uganda                                      6283 non-null
uint8
 78  Country_United Arab Emirates                        6283 non-null
uint8
 79  Country_United Kingdom                              6283 non-null
uint8
 80  Country_United States                               6283 non-null
uint8
 81  Country_Vietnam                                     6283 non-null
uint8
 82  Country_unknown                                     6283 non-null
uint8
 83  LastNotableActivity_Approached upfront              6283 non-null
uint8
 84  LastNotableActivity_Email Bounced                   6283 non-null
uint8
 85  LastNotableActivity_Email Link Clicked              6283 non-null
uint8
 86  LastNotableActivity_Email Marked Spam               6283 non-null
uint8
 87  LastNotableActivity_Email Opened                    6283 non-null
uint8
 88  LastNotableActivity_Form Submitted on Website       6283 non-null
uint8
 89  LastNotableActivity_Had a Phone Conversation        6283 non-null
uint8
 90  LastNotableActivity_Modified                        6283 non-null
uint8
 91  LastNotableActivity_Olark Chat Conversation         6283 non-null
uint8
 92  LastNotableActivity_Page Visited on Website         6283 non-null
uint8
 93  LastNotableActivity_Resubscribed to emails          6283 non-null
uint8
 94  LastNotableActivity_SMS Sent                        6283 non-null
uint8
 95  LastNotableActivity_Unreachable                     6283 non-null
uint8
 96  LastNotableActivity_Unsubscribed                    6283 non-null
uint8
 97  LastNotableActivity_View in browser link Clicked    6283 non-null
uint8
dtypes: float64(2), int64(2), uint8(94)
memory usage: 822.2 KB
```

Scale Features

```
#scaling numeric columns
scaler = StandardScaler()
num_cols=X_train.select_dtypes(include=['float64', 'int64']).columns
X_train[num_cols] = scaler.fit_transform(X_train[num_cols])
X_train.head()
```

```
      TotalVisits  TotalTimeSpentOnWebsite  PageViewsPerVisit  \
5555     -0.74529                -0.167824          -0.671585
568      -0.04857                 1.604517          -0.414920
3810     -0.39693                 1.788944          -0.158254
903       0.29979                -0.863113           0.868407
1831      4.48011                 1.921731           0.185677

      FreeCopyOfMasteringInterview  LeadOrigin_Landing Page Submission
\
5555                       -0.66636                                   0

568                         1.50069                                   1

3810                       -0.66636                                   0

903                        -0.66636                                   1

1831                        1.50069                                   1


      LeadOrigin_Lead Add Form  LeadOrigin_Lead Import  \
5555                         0                       0
568                          0                       0
3810                         0                       0
903                          0                       0
1831                         0                       0

      CurrentOccupation_Housewife  CurrentOccupation_Other  \
5555                            0                       0
568                             0                       0
3810                            0                       0
903                             0                       0
1831                            0                       0

      CurrentOccupation_Student  ...  \
5555                          0  ...
568                           0  ...
3810                          0  ...
903                           0  ...
1831                          0  ...

      LastNotableActivity_Form Submitted on Website  \
5555                                              0
```

```
568                                                    0
3810                                                   0
903                                                    0
1831                                                   0

      LastNotableActivity_Had a Phone Conversation  \
5555                                                  0
568                                                   0
3810                                                  0
903                                                   0
1831                                                  0

      LastNotableActivity_Modified  \
5555                              1
568                               0
3810                              0
903                               1
1831                              1

      LastNotableActivity_Olark Chat Conversation  \
5555                                            0
568                                             0
3810                                            0
903                                             0
1831                                            0

      LastNotableActivity_Page Visited on Website  \
5555                                            0
568                                             0
3810                                            0
903                                             0
1831                                            0

      LastNotableActivity_Resubscribed to emails  \
5555                                           0
568                                            0
3810                                           0
903                                            0
1831                                           0

      LastNotableActivity_SMS Sent  LastNotableActivity_Unreachable  \
5555                              0                                 0
568                               0                                 0
3810                              0                                 0
903                               0                                 0
1831                              0                                 0

      LastNotableActivity_Unsubscribed  \
5555                                 0
568                                  0
```

```
3810                                        0
903                                         0
1831                                        0

      LastNotableActivity_View in browser link Clicked
5555                                                  0
568                                                   0
3810                                                  0
903                                                   0
1831                                                  0

[5 rows x 98 columns]
```

Model Building using Stats Model & RFE

```
logreg = LogisticRegression()
rfe = RFE(logreg, n_features_to_select=15, step=1)
rfe = rfe.fit(X_train, y_train)
rfe.support_
```

```
array([False,  True, False, False, False,  True, False,  True, False,
       False, False,  True, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False,  True, False, False, False,
        True,  True, False,  True, False, False, False,  True, False,
        True, False, False, False, False, False, False, False, False,
       False, False, False, False, False,  True, False, False, False,
       False, False,  True, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False,  True,
        True, False, False, False, False,  True, False, False])
```

```
list(zip(X_train.columns, rfe.support_, rfe.ranking_))
```

```
[('TotalVisits', False, 53),
 ('TotalTimeSpentOnWebsite', True, 1),
 ('PageViewsPerVisit', False, 54),
 ('FreeCopyOfMasteringInterview', False, 73),
 ('LeadOrigin_Landing Page Submission', False, 12),
 ('LeadOrigin_Lead Add Form', True, 1),
 ('LeadOrigin_Lead Import', False, 72),
 ('CurrentOccupation_Housewife', True, 1),
 ('CurrentOccupation_Other', False, 10),
 ('CurrentOccupation_Student', False, 9),
 ('CurrentOccupation_Unemployed', False, 8),
 ('CurrentOccupation_Working Professional', True, 1),
 ('Specialization_Banking, Investment And Insurance', False, 15),
 ('Specialization_Business Administration', False, 20),
 ('Specialization_E-Business', False, 19),
 ('Specialization_E-COMMERCE', False, 18),
```

```
('Specialization_Finance Management', False, 21),
('Specialization_Healthcare Management', False, 25),
('Specialization_Hospitality Management', False, 74),
('Specialization_Human Resource Management', False, 23),
('Specialization_IT Projects Management', False, 17),
('Specialization_International Business', False, 27),
('Specialization_Marketing Management', False, 14),
('Specialization_Media and Advertising', False, 26),
('Specialization_Operations Management', False, 24),
('Specialization_Retail Management', False, 28),
('Specialization_Rural and Agribusiness', False, 13),
('Specialization_Services Excellence', False, 29),
('Specialization_Supply Chain Management', False, 16),
('Specialization_Travel and Tourism', False, 22),
('LeadSource_Direct Traffic', False, 38),
('LeadSource_Google', False, 49),
('LeadSource_Olark Chat', True, 1),
('LeadSource_Organic Search', False, 48),
('LeadSource_Reference', False, 62),
('LeadSource_Referral Sites', False, 42),
('LeadSource_Social Media', True, 1),
('LeadSource_Welingak Website', True, 1),
('LastActivity_Converted to Lead', False, 39),
('LastActivity_Email Bounced', True, 1),
('LastActivity_Email Link Clicked', False, 70),
('LastActivity_Email Opened', False, 7),
('LastActivity_Form Submitted on Website', False, 44),
('LastActivity_Olark Chat Conversation', True, 1),
('LastActivity_Page Visited on Website', False, 50),
('LastActivity_SMS Sent', True, 1),
('Country_Asia/Pacific Region', False, 77),
('Country_Australia', False, 71),
('Country_Bahrain', False, 46),
('Country_Bangladesh', False, 76),
('Country_Belgium', False, 68),
('Country_Canada', False, 37),
('Country_China', False, 78),
('Country_Denmark', False, 83),
('Country_France', False, 34),
('Country_Germany', False, 36),
('Country_Ghana', False, 64),
('Country_Hong Kong', False, 59),
('Country_India', False, 32),
('Country_Italy', True, 1),
('Country_Kenya', False, 82),
('Country_Kuwait', False, 35),
('Country_Liberia', False, 80),
('Country_Malaysia', False, 79),
('Country_Netherlands', False, 58),
```

```
 ('Country_Nigeria', True, 1),
 ('Country_Oman', False, 30),
 ('Country_Philippines', False, 65),
 ('Country_Qatar', False, 69),
 ('Country_Russia', False, 84),
 ('Country_Saudi Arabia', False, 63),
 ('Country_Singapore', False, 56),
 ('Country_South Africa', False, 33),
 ('Country_Sri Lanka', False, 81),
 ('Country_Sweden', False, 51),
 ('Country_Switzerland', False, 61),
 ('Country_Tanzania', False, 75),
 ('Country_Uganda', False, 45),
 ('Country_United Arab Emirates', False, 31),
 ('Country_United Kingdom', False, 55),
 ('Country_United States', False, 66),
 ('Country_Vietnam', False, 60),
 ('Country_unknown', False, 43),
 ('LastNotableActivity_Approached upfront', False, 47),
 ('LastNotableActivity_Email Bounced', False, 41),
 ('LastNotableActivity_Email Link Clicked', False, 4),
 ('LastNotableActivity_Email Marked Spam', False, 57),
 ('LastNotableActivity_Email Opened', False, 6),
 ('LastNotableActivity_Form Submitted on Website', False, 67),
 ('LastNotableActivity_Had a Phone Conversation', True, 1),
 ('LastNotableActivity_Modified', True, 1),
 ('LastNotableActivity_Olark Chat Conversation', False, 2),
 ('LastNotableActivity_Page Visited on Website', False, 5),
 ('LastNotableActivity_Resubscribed to emails', False, 11),
 ('LastNotableActivity_SMS Sent', False, 40),
 ('LastNotableActivity_Unreachable', True, 1),
 ('LastNotableActivity_Unsubscribed', False, 3),
 ('LastNotableActivity_View in browser link Clicked', False, 52)]
```

```
#list of RFE supported columns
rfe_columns = X_train.columns[rfe.support_]
rfe_columns
```

```
Index(['TotalTimeSpentOnWebsite', 'LeadOrigin_Lead Add Form',
       'CurrentOccupation_Housewife', 'CurrentOccupation_Working
Professional',
       'LeadSource_Olark Chat', 'LeadSource_Social Media',
       'LeadSource_Welingak Website', 'LastActivity_Email Bounced',
       'LastActivity_Olark Chat Conversation', 'LastActivity_SMS
Sent',
       'Country_Italy', 'Country_Nigeria',
       'LastNotableActivity_Had a Phone Conversation',
       'LastNotableActivity_Modified',
'LastNotableActivity_Unreachable'],
      dtype='object')
```

```
X_train.columns[~rfe.support_]
```

```
Index(['TotalVisits', 'PageViewsPerVisit',
'FreeCopyOfMasteringInterview',
       'LeadOrigin_Landing Page Submission', 'LeadOrigin_Lead Import',
       'CurrentOccupation_Other', 'CurrentOccupation_Student',
       'CurrentOccupation_Unemployed',
       'Specialization_Banking, Investment And Insurance',
       'Specialization_Business Administration', 'Specialization_E-
Business',
       'Specialization_E-COMMERCE', 'Specialization_Finance
Management',
       'Specialization_Healthcare Management',
       'Specialization_Hospitality Management',
       'Specialization_Human Resource Management',
       'Specialization_IT Projects Management',
       'Specialization_International Business',
       'Specialization_Marketing Management',
       'Specialization_Media and Advertising',
       'Specialization_Operations Management',
       'Specialization_Retail Management',
       'Specialization_Rural and Agribusiness',
       'Specialization_Services Excellence',
       'Specialization_Supply Chain Management',
       'Specialization_Travel and Tourism', 'LeadSource_Direct
Traffic',
       'LeadSource_Google', 'LeadSource_Organic Search',
       'LeadSource_Reference', 'LeadSource_Referral Sites',
       'LastActivity_Converted to Lead', 'LastActivity_Email Link
Clicked',
       'LastActivity_Email Opened', 'LastActivity_Form Submitted on
Website',
       'LastActivity_Page Visited on Website', 'Country_Asia/Pacific
Region',
       'Country_Australia', 'Country_Bahrain', 'Country_Bangladesh',
       'Country_Belgium', 'Country_Canada', 'Country_China',
'Country_Denmark',
       'Country_France', 'Country_Germany', 'Country_Ghana',
       'Country_Hong Kong', 'Country_India', 'Country_Kenya',
'Country_Kuwait',
       'Country_Liberia', 'Country_Malaysia', 'Country_Netherlands',
       'Country_Oman', 'Country_Philippines', 'Country_Qatar',
       'Country_Russia', 'Country_Saudi Arabia', 'Country_Singapore',
       'Country_South Africa', 'Country_Sri Lanka', 'Country_Sweden',
       'Country_Switzerland', 'Country_Tanzania', 'Country_Uganda',
       'Country_United Arab Emirates', 'Country_United Kingdom',
       'Country_United States', 'Country_Vietnam', 'Country_unknown',
       'LastNotableActivity_Approached upfront',
       'LastNotableActivity_Email Bounced',
       'LastNotableActivity_Email Link Clicked',
```

```
       'LastNotableActivity_Email Marked Spam',
       'LastNotableActivity_Email Opened',
       'LastNotableActivity_Form Submitted on Website',
       'LastNotableActivity_Olark Chat Conversation',
       'LastNotableActivity_Page Visited on Website',
       'LastNotableActivity_Resubscribed to emails',
       'LastNotableActivity_SMS Sent',
'LastNotableActivity_Unsubscribed',
       'LastNotableActivity_View in browser link Clicked'],
      dtype='object')
```

```
X_train_sm = sm.add_constant(X_train[rfe_columns])
logm1 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm1.fit()
res.summary()
```

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results

==============================================================================
========
Dep. Variable:              Converted   No. Observations:
6283
Model:                            GLM   Df Residuals:
6267
Model Family:                Binomial   Df Model:
15
Link Function:                   logit   Scale:
1.0000
Method:                           IRLS   Log-Likelihood:
-2659.2
Date:                Wed, 01 Nov 2023   Deviance:
5318.4
Time:                        00:19:41   Pearson chi2:
6.64e+03
No. Iterations:                     22

Covariance Type:            nonrobust

==============================================================================
==========================================
                                                    coef    std err
z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------------
const                                          -1.2021      0.056
-21.302      0.000      -1.313      -1.091
TotalTimeSpentOnWebsite                         1.0985      0.040
27.776      0.000       1.021       1.176
```

```
LeadOrigin_Lead Add Form                          4.0696      0.232
17.532      0.000       3.615       4.525
CurrentOccupation_Housewife                      24.0180   2.53e+04
0.001      0.999   -4.96e+04    4.96e+04
CurrentOccupation_Working Professional            2.8015      0.199
14.050      0.000       2.411       3.192
LeadSource_Olark Chat                             1.1442      0.104
11.030      0.000       0.941       1.347
LeadSource_Social Media                           1.2430      0.480
2.589      0.010       0.302       2.184
LeadSource_Welingak Website                       1.5746      0.755
2.087      0.037       0.096       3.053
LastActivity_Email Bounced                       -1.6301      0.340
-4.789      0.000      -2.297      -0.963
LastActivity_Olark Chat Conversation             -1.1767      0.177
-6.646      0.000      -1.524      -0.830
LastActivity_SMS Sent                             1.1618      0.073
15.868      0.000       1.018       1.305
Country_Italy                                   -24.8247   4.48e+04
-0.001      1.000   -8.78e+04    8.77e+04
Country_Nigeria                                 -23.0950   3.92e+04
-0.001      1.000   -7.68e+04    7.67e+04
LastNotableActivity_Had a Phone Conversation     23.7086   3.08e+04
0.001      0.999   -6.03e+04    6.03e+04
LastNotableActivity_Modified                     -0.8694      0.080
-10.850      0.000      -1.026      -0.712
LastNotableActivity_Unreachable                   2.2306      0.599
3.727      0.000       1.058       3.404
==============================================================================
==========================================
"""
```

p-value of CurrentOccupation_Housewife, LastNotableActivity_Had a Phone Conversation is very high => Let's drop them

```
rfe_columns = rfe_columns.drop(['CurrentOccupation_Housewife',
'LastNotableActivity_Had a Phone Conversation'], 1)

X_train_sm = sm.add_constant(X_train[rfe_columns])
logm2 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm2.fit()
res.summary()

<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results

==============================================================================
========
Dep. Variable:              Converted   No. Observations:
```

```
6283
Model:                            GLM   Df Residuals:
6269
Model Family:                Binomial   Df Model:
13
Link Function:                   logit   Scale:
1.0000
Method:                           IRLS   Log-Likelihood:
-2667.5
Date:                Wed, 01 Nov 2023   Deviance:
5335.0
Time:                        00:19:53   Pearson chi2:
6.66e+03
No. Iterations:                     21

Covariance Type:            nonrobust

==============================================================================
====================================
                                             coef    std err
z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------
const                                     -1.1914      0.056    -
21.184      0.000      -1.302      -1.081
TotalTimeSpentOnWebsite                    1.0966      0.039
27.783      0.000       1.019       1.174
LeadOrigin_Lead Add Form                   4.0846      0.232
17.624      0.000       3.630       4.539
CurrentOccupation_Working Professional     2.7999      0.199
14.056      0.000       2.409       3.190
LeadSource_Olark Chat                      1.1359      0.104
10.966      0.000       0.933       1.339
LeadSource_Social Media                    1.2333      0.480
2.570      0.010       0.293       2.174
LeadSource_Welingak Website                1.5515      0.754
2.056      0.040       0.073       3.030
LastActivity_Email Bounced                -1.6378      0.340     -
4.813      0.000      -2.305      -0.971
LastActivity_Olark Chat Conversation      -1.1817      0.177     -
6.677      0.000      -1.529      -0.835
LastActivity_SMS Sent                      1.1525      0.073
15.763      0.000       1.009       1.296
Country_Italy                            -23.8213   2.71e+04     -
0.001      0.999   -5.32e+04    5.32e+04
Country_Nigeria                          -22.0980   2.38e+04     -
0.001      0.999   -4.66e+04    4.66e+04
LastNotableActivity_Modified              -0.8704      0.080    -
10.882      0.000      -1.027      -0.714
```

```
LastNotableActivity_Unreachable            2.2206      0.598
3.711      0.000       1.048       3.394
==============================================================================
==================================
"""
```

p-value of Country_Italy, Country_Nigeria is very high => Let's drop them

```
rfe_columns = rfe_columns.drop(['Country_Italy', 'Country_Nigeria'],
1)

X_train_sm = sm.add_constant(X_train[rfe_columns])
logm2 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm2.fit()
res.summary()

<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results

==============================================================================
========
Dep. Variable:               Converted   No. Observations:
6283
Model:                             GLM   Df Residuals:
6271
Model Family:                Binomial   Df Model:
11
Link Function:                   logit   Scale:
1.0000
Method:                           IRLS   Log-Likelihood:
-2671.8
Date:                Wed, 01 Nov 2023   Deviance:
5343.6
Time:                        00:20:44   Pearson chi2:
6.66e+03
No. Iterations:                     7

Covariance Type:            nonrobust

==============================================================================
==================================
                                            coef    std err
z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
----------------------------------
const                                    -1.1924      0.056    -
21.217      0.000      -1.303      -1.082
TotalTimeSpentOnWebsite                   1.0943      0.039
27.769      0.000       1.017       1.172
```

```
LeadOrigin_Lead Add Form                    4.0838      0.232
17.626      0.000       3.630       4.538
CurrentOccupation_Working Professional     2.8012      0.199
14.070      0.000       2.411       3.191
LeadSource_Olark Chat                       1.1362      0.104
10.976      0.000       0.933       1.339
LeadSource_Social Media                     1.2326      0.480
2.568       0.010       0.292       2.173
LeadSource_Welingak Website                 1.5525      0.754
2.058       0.040       0.074       3.031
LastActivity_Email Bounced                 -1.6380      0.340     -
4.815       0.000      -2.305      -0.971
LastActivity_Olark Chat Conversation       -1.1836      0.177     -
6.690       0.000      -1.530      -0.837
LastActivity_SMS Sent                       1.1468      0.073
15.703      0.000       1.004       1.290
LastNotableActivity_Modified               -0.8670      0.080    -
10.845      0.000      -1.024      -0.710
LastNotableActivity_Unreachable             2.2211      0.598
3.712       0.000       1.048       3.394
==============================================================================
==================================
"""
```

All' the p-values are less we can check the Variance Inflation Factor to see if there is any correlation between the variables

```
# Check for the VIF values of the feature variables.
from statsmodels.stats.outliers_influence import
variance_inflation_factor

# Create a dataframe that will contain the names of all the feature
variables and their respective VIFs
vif = pd.DataFrame()
vif['Features'] = X_train[rfe_columns].columns
vif['VIF'] = [variance_inflation_factor(X_train[rfe_columns].values,
i) for i in range(X_train[rfe_columns].shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
```

```
                                  Features   VIF
3                    LeadSource_Olark Chat  1.60
7     LastActivity_Olark Chat Conversation  1.58
1                 LeadOrigin_Lead Add Form  1.53
9             LastNotableActivity_Modified  1.45
5              LeadSource_Welingak Website  1.32
0                  TotalTimeSpentOnWebsite  1.28
8                    LastActivity_SMS Sent  1.24
2   CurrentOccupation_Working Professional  1.15
```

```
6                LastActivity_Email Bounced  1.09
4                   LeadSource_Social Media  1.01
10         LastNotableActivity_Unreachable  1.00
```

VIF Looks good

Getting Predictions

```
# Getting the Predicted values on the train set
y_train_pred = res.predict(X_train_sm)
y_train_pred[:10]
```

```
5555    0.095955
568     0.637238
3810    0.682484
903     0.047253
1831    0.510886
4530    0.112623
8673    0.386554
3319    0.394920
8171    0.245133
7650    0.294198
dtype: float64
```

```
y_train_pred = y_train_pred.values.reshape(-1)
y_train_pred[:10]
```

```
array([0.09595461, 0.6372383 , 0.68248419, 0.04725252, 0.51088592,
       0.11262307, 0.3865545 , 0.39492023, 0.24513301, 0.29419792])
```

```
y_train_pred_final = pd.DataFrame({'Converted':y_train.values,
'Converted_prob':y_train_pred})
y_train_pred_final['Prospect ID'] = y_train.index
y_train_pred_final.head()
```

```
   Converted  Converted_prob  Prospect ID
0          0        0.095955         5555
1          1        0.637238          568
2          1        0.682484         3810
3          0        0.047253          903
4          1        0.510886         1831
```

```
y_train_pred_final['Predicted'] =
y_train_pred_final.Converted_prob.map(lambda x: 1 if x > 0.5 else 0)

# Let's see the head
y_train_pred_final.head()
```

```
   Converted  Converted_prob  Prospect ID  Predicted
0          0        0.095955         5555          0
1          1        0.637238          568          1
```

```
2          1        0.682484         3810          1
3          0        0.047253          903          0
4          1        0.510886         1831          1
```

Checking Confusion matrix

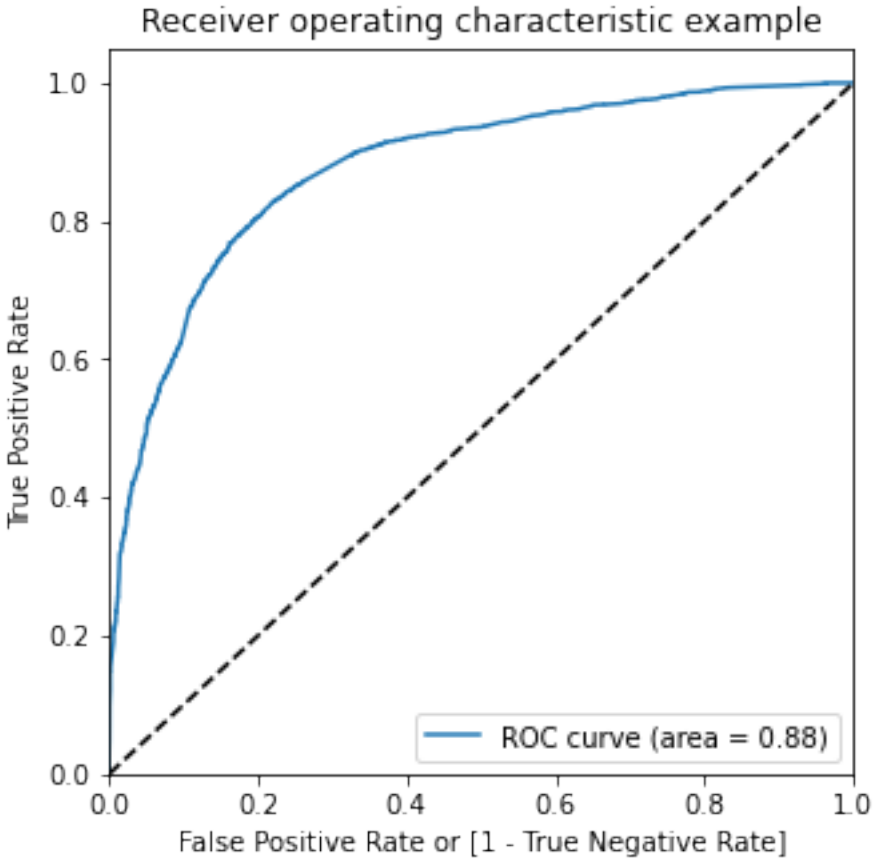
```
from sklearn import metrics
confusion = metrics.confusion_matrix(y_train_pred_final.Converted,
y_train_pred_final.Predicted )
print('Confusion', confusion)

# overall accuracy
print('Overall Accuracy: ',
metrics.accuracy_score(y_train_pred_final.Converted,
y_train_pred_final.Predicted))

TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives

print('Sensitivity', TP / float(TP+FN))
print('Specificity', TN / float(TN+FP))
```

```
Confusion [[3440  448]
 [ 751 1644]]
Overall Accuracy:  0.8091675950978832
Sensitivity 0.6864300626304801
Specificity 0.8847736625514403
```

ROC CURVE

```
def draw_roc( actual, probs ):
    fpr, tpr, thresholds = metrics.roc_curve( actual, probs,
                                              drop_intermediate =
False )
    auc_score = metrics.roc_auc_score( actual, probs )
    plt.figure(figsize=(5, 5))
    plt.plot( fpr, tpr, label='ROC curve (area = %0.2f)' % auc_score )
    plt.plot([0, 1], [0, 1], 'k--')
    plt.xlim([0.0, 1.0])
    plt.ylim([0.0, 1.05])
    plt.xlabel('False Positive Rate or [1 - True Negative Rate]')
    plt.ylabel('True Positive Rate')
    plt.title('Receiver operating characteristic example')
    plt.legend(loc="lower right")
    plt.show()

    return None
```

```
fpr, tpr, thresholds =
metrics.roc_curve( y_train_pred_final.Converted,
y_train_pred_final.Converted_prob, drop_intermediate = False )

draw_roc(y_train_pred_final.Converted,
y_train_pred_final.Converted_prob)
```



ROC Curve is close to the left hand and near on the top => It's good

Making Predictions

```
#scaling test set

num_cols=X_test.select_dtypes(include=['float64', 'int64']).columns
X_test[num_cols] = scaler.fit_transform(X_test[num_cols])
X_test.head()
```

	TotalVisits	TotalTimeSpentOnWebsite	PageViewsPerVisit	\
4190	-0.728622	-0.313269	-0.652297	
5186	0.590047	1.065190	0.139170	
7032	-0.398955	1.553204	-0.124652	
5977	-1.058290	-0.865017	-1.179942	
7795	-0.398955	-0.839523	-0.652297	

```
      FreeCopyOfMasteringInterview  LeadOrigin_Landing Page Submission
\
4190                      1.435349                                   1

5186                      1.435349                                   1

7032                      1.435349                                   1

5977                     -0.696695                                   0

7795                     -0.696695                                   0


      LeadOrigin_Lead Add Form  LeadOrigin_Lead Import  \
4190                         0                       0
5186                         0                       0
7032                         0                       0
5977                         0                       0
7795                         0                       0

      CurrentOccupation_Housewife  CurrentOccupation_Other  \
4190                            0                       0
5186                            0                       0
7032                            0                       0
5977                            0                       0
7795                            0                       0

      CurrentOccupation_Student  ...  \
4190                          0  ...
5186                          0  ...
7032                          0  ...
5977                          0  ...
7795                          0  ...

      LastNotableActivity_Form Submitted on Website  \
4190                                              0
5186                                              0
7032                                              0
5977                                              0
7795                                              0

      LastNotableActivity_Had a Phone Conversation  \
4190                                             0
5186                                             0
7032                                             0
5977                                             0
7795                                             0

      LastNotableActivity_Modified  \
```

```
4190                                1
5186                                0
7032                                1
5977                                1
7795                                1

      LastNotableActivity_Olark Chat Conversation  \
4190                                             0
5186                                             0
7032                                             0
5977                                             0
7795                                             0

      LastNotableActivity_Page Visited on Website  \
4190                                             0
5186                                             0
7032                                             0
5977                                             0
7795                                             0

      LastNotableActivity_Resubscribed to emails  \
4190                                            0
5186                                            0
7032                                            0
5977                                            0
7795                                            0

      LastNotableActivity_SMS Sent  LastNotableActivity_Unreachable  \
4190                             0                                 0
5186                             0                                 0
7032                             0                                 0
5977                             0                                 0
7795                             0                                 0

      LastNotableActivity_Unsubscribed  \
4190                                  0
5186                                  0
7032                                  0
5977                                  0
7795                                  0

      LastNotableActivity_View in browser link Clicked
4190                                                 0
5186                                                 0
7032                                                 0
5977                                                 0
7795                                                 0

[5 rows x 98 columns]
```

```
X_test = X_test[rfe_columns]
X_test.head()
```

```
      TotalTimeSpentOnWebsite  LeadOrigin_Lead Add Form  \
4190                -0.313269                         0
5186                 1.065190                         0
7032                 1.553204                         0
5977                -0.865017                         0
7795                -0.839523                         0

      CurrentOccupation_Working Professional  LeadSource_Olark Chat  \
4190                                       0                      0
5186                                       0                      0
7032                                       0                      0
5977                                       0                      1
7795                                       0                      0

      LeadSource_Social Media  LeadSource_Welingak Website  \
4190                        0                            0
5186                        0                            0
7032                        0                            0
5977                        0                            0
7795                        0                            0

      LastActivity_Email Bounced  LastActivity_Olark Chat Conversation
\
4190                           0                                      0

5186                           0                                      0

7032                           0                                      0

5977                           0                                      1

7795                           0                                      0


      LastActivity_SMS Sent  LastNotableActivity_Modified  \
4190                      0                             1
5186                      0                             0
7032                      1                             1
5977                      0                             1
7795                      1                             1

      LastNotableActivity_Unreachable
4190                                0
5186                                0
7032                                0
5977                                0
7795                                0
```

```
X_test_sm = sm.add_constant(X_test)

y_test_pred = res.predict(X_test_sm)

y_test_pred[:10]
```

```
4190    0.083008
5186    0.493306
7032    0.687195
5977    0.045075
7795    0.138087
6457    0.866202
7214    0.991389
5107    0.268402
2635    0.051745
6785    0.770134
dtype: float64
```

```
# Converting y_pred to a dataframe which is an array
y_pred_1 = pd.DataFrame(y_test_pred)

# Let's see the head
y_pred_1.head()
```

	0
4190	0.083008
5186	0.493306
7032	0.687195
5977	0.045075
7795	0.138087

```
# Converting y_test to dataframe
y_test_df = pd.DataFrame(y_test)
# Putting CustID to index
y_test_df['Prospect ID'] = y_test_df.index
# Removing index for both dataframes to append them side by side
y_pred_1.reset_index(drop=True, inplace=True)
y_test_df.reset_index(drop=True, inplace=True)
# Appending y_test_df and y_pred_1
y_pred_final = pd.concat([y_test_df, y_pred_1],axis=1)
y_pred_final.head()
```

	Converted	Prospect ID	0
0	0	4190	0.083008
1	1	5186	0.493306
2	0	7032	0.687195
3	0	5977	0.045075
4	1	7795	0.138087